Global Journal of Advanced Computer Science and Technology

Volume No. 12

Issue No. 1

January - April 2024

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

Global Journal of Advanced Computer Science and Technology

Aims and Scope

This journal Global Journal of Advanced Computer Science & Technology deals with information technology, its evolution and future prospects and its relationship with the Business Management. It addresses technological, managerial, political, economic and organizational aspects of the application of IT in relationship with Business Management. The journal will serve as a comprehensiveresource for policy makers, government officials, academicians, and practitioners. GJACST promotes and coordinates the developments in the IT based applications of business management and presents thestrategic roles of IT and management towards sustainable development.

Global Journal of Advanced Computer Science and Technology

Global Journal of Advanced Computer Science and Technology

(Volume No. 12, Issue No. 1, January - April 2024)

Contents

Sr. No	Article/ Autors	Pg No
01	Metadata Are Everywhere: Use And Problems Of Metadata In Trisquel GNU/Linux - *Daniel Martínez-Ávila*	01-20
02	Design and Analysis of a Big Data Technologies and WI [Web Intelligence] and Using Hadoop: A Tool for Solution - *Jitender Rai*	21-29
03	Users' Information Search Strategies in Computerised Libraries of Degree College Sivasagar District, Assam - *Hima Devi, Jnyandeep Saikia*	30-37
04	STEGANOGRAPHY- Art Of Hiding Information - *Geeta Wahi*	38-43
05	Development Of Predictive Simulation Models For Drug Dissolution Parameters Computing - *JDr. H. B. Bhadka*	44-53

Metadata Are Everywhere: Use And Problems Of Metadata In Trisquel GNU/Linux

Daniel Martínez-Ávila
Information Science Department, São Paulo State University, Marília-SP, Brazil. E-mail: dmartinezavila@marilia.unesp.br

ABSTRACT

The term metadata has different meanings for different communities. The nuances of those meanings can also be traced back in history and analyzed according the technological milestones of our age. In the current state of technology, most approaches to metadata must be based on semantic statements. Standardization is another key aspect of the current metadata approach, especially to deal with the mapping of different schemes and uses. However, there are several cases of misuses of metadata that are being detrimental to the growth and development of those technologies. As an example, the use of the application of metadata in the Trisquel GNU/Linux operating system has been analyzed to illustrate some of these problems.

Keywords—Metadata, Semantics, Standardization, Trisquel GNU/Linux.

INTRODUCTION

The term metadata has different meanings for different communities. The nuances of those meanings for the different communities can also be traced back in history and analyzed according the technological milestones of our age. For instance, Anne Gilliland (2008) highlights that for the geospatial data community, until the mid-1990s, metadata was a suite of industry or disciplinary standards, the internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system; for an Internet resource provider metadata is the information encoded in HTML meta tags for the purposes of making a Web site easier to find; in the digitization of images metadata primarily is the information in the header field about the image file, the imaging process, and image rights; for a social science data archivist metadata are the systems and research documentation necessary to run and interpret a magnetic tape containing raw research data; for an electronic records archivist it is the contextual, processing, preservation, and use information needed to identify and document the scope, authenticity, and integrity of an active or archival record in an electronic record-keeping or archival

preservation system; and so on. In attempting to provide definitions that can be used for different communities, the National Information Standards Organization provided a more inclusive definition: "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information" (National Information Standards Organization 2004, 1); while, trying to provide a more inclusive definitions for the future and potential needs of an electronic environment, other authors have talked about "The sum total of what one can say about any information object at any level of aggregation" (Gilliland 2008, p.2). In this vein, when talking about information objects, one cannot help thinking of concepts such object oriented programming, classes and the like in computer science. Following with Gilliland's definition, it is stated that information object is anything that can be addressed and manipulated as a discrete entity by a human being or an information system. The object may comprise a single item, an aggregate of many items, or the entire database or record-keeping system. Recapiltulating, and going back to an inclusive definition that can be valid for both Library and Information Science at their broadest meanings, the most common definition of metadata is "data about data", or in other words, information about information resources where it can can describe resources at any level of aggregation; it can describe a collection, a single resource, or a component part of a larger resource (for example, a photograph in an article); it can describe works, expressions, manifestations, or items; and it can be embedded in the information resource or it can be stored separately (e.g. in a database).

Semantic approach to metadata in electronic and computer contexts

In the context of the Semantic Web, and in a document published by the American Library Association, Karen Coyle states that metadata are constructued, constructive and actionable (Coyle 2010, p.6). She says that they are constructed because Metadata are not found in nature. It is entirely an invention; it is an artificiality. Much more as the abstract objects that are constructed in programs. Metadata are constructive because they are constructed for some purpose, some activity, to solve some problem. Same way an algorithm is a set of instructions to solve a problem and programs are also written for the same reason, metadata are (should be) constructive also in a pragmatic way to achieve a goal in a given context. Coyle also points out that the proliferation of metadata formats that seem similar on the surface is often evidence of different definitions of needs or of different contexts. As in the case of ontologies, one may dream of a universal set of metadata for some set of things, like biological entities, printed books, or a calendar of events, and be likely to be disappointed in practice. Finally, metadata are actionable because the point of metadata is to be useful in some way. This means that it is important that one can act on the metadata in a way that satisfies some needs. This aspect is key for the concept of information objects, for instance, in the context of an operating system and web applications. Related

to the semantic context, metadata consists of statements we make about resources to help us find, identify, use, manage, evaluate, and preserve them. And in this vein, there has been aswers from three traditions: the database management systems (with relational database schemas), the library cataloging traditions (with standards such as MARC and AACR2), and the World Wide Web (since the mid-1990's). Trying to unify these approaches it can be stated that the basic unit of metadata is a statement, a statement consists of a property (element) and a value (i.e., a property/value pair), metadata statements describe resources, and each statement is a logically distinct semantic unit (see Figure 1).

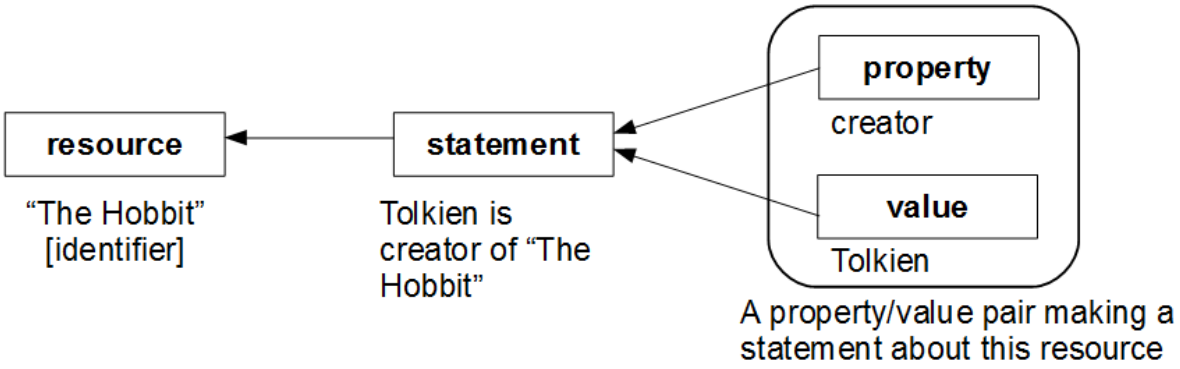


Figure 1: Example of metadata statements. It is stated that Tolkien is the creator of "The Hobbit"

Each "statement" is a discrete piece of human-readable and machine-readable/computer processable information. It is actionable and can be used by programs and other agents. Each statement is a logical unit that a computer can process and that could be taken out of the context (for instance a bibliographic record) and stand on its own. The value of each statement is the exact character string; a computer must process this character string exactly as it is. This is a semantic rather than a record based view of metadata/bibliographic data and it is the foundational concept underlying Linked Data and the Semantic Web. It should be noted that some fields or elements in their public display might contain more than one value, such as multiple names or subjects in a bibliographic record. However, each needs to be taken as a separate logical statement, understanding that the property applies to each value separately. For example: a resource on eastern religions has three different subjects separated by semicolon: Buddhism; Hinduism; Shintoism. This description consists and should be treated as three semantically separate statements: Buddhism is Subject of ResourceXYZ (ResourceXYZ hasSubject Buddhism), Hinduism is Subject of ResourceXYZ (ResourceXYZ hasSubject Hinduism), Shintoism is Subject of ResourceXYZ (ResourceXYZ hasSubject Shintoism). In the following example (see Figure 2), the bibliographic record of the book "The girl with the dragon tattoo" lists four different subjects (values): "Missing persons --Fiction," "Rich people --Fiction," "Corruption --Fiction," and "Stockholm (Sweden)

--Fiction." However, although the property "Subject" is stated only once in the display, the record contains four logically and semantically distinct statements about the Subject of the resource. This view is vital for the requirements of Linked Data and the Semantic Web since each statement must be treated as an independent logical unit. Figure 3 shows a linked data example/semantic statements about the same resource where each statement consists of a property and a value pair.

LC control no.: 2009291782
LCCN permalink: http://lccn.loc.gov/2009291782
Type of material: Book (Print, Microform, Electronic, etc.)
Personal name: Larsson, Stieg, 1954-2004.
Uniform title: Män som hatar kvinnor. English
Main title: The girl with the dragon tattoo / Stieg Larsson ; translated from the Swedish by Reg Keeland.
Edition: 1st Vintage Crime/Black Lizard ed.
Published/Created: New York : Vintage Crime/Black Lizard, 2009.
Description: 590, 10 p. ; 21 cm.
ISBN: 9780307454546 (pbk.)
Subjects: Missing persons --Fiction.
Rich people --Fiction.
Corruption --Fiction.
Stockholm (Sweden) --Fiction.
Form/Genre: Suspense fiction.

Figure 2: Traditional metadata record example (library catalog record)

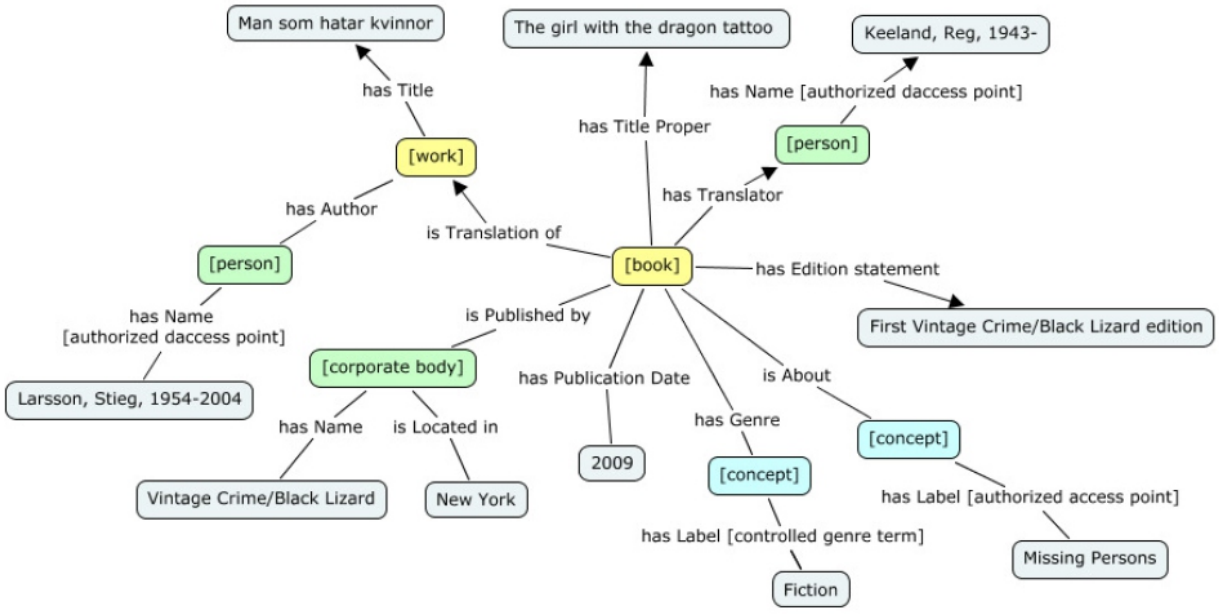


Figure 3: Linked data example

Metadata are everywhere (especially in electronic contexts)

In electronic contexts, metadata generation is as important as relatively easy to generate. It can be stated that in order to be used and processed by programs in an effective way, actionable, there must be some information attached to the objects carrying selected elements that represent some of the connotation context and interpretation to make the information (not just the data) processable by the programs. For example, Figure 4 shows descriptive metadata embedded in the "properties" of a Microsoft Word document. It should be noted too that some of the statements, especially at the bottom, corresponds with administrative metadata. Figure 5 is an example of metadata automatically generated by a digital camera for a digital photograph, also displayed in Microsoft Windows. In this case, the statements correspond mostly with technical metadata. These metadata can be understood and used by the system in huge variety of ways, from arranging and filtering searches to manipulating the object using appropriate applications. Going back to the bibliographic use of metadata in library and online catalogs, Figure 6 shows and example of descriptive metadata as displayed to users of a library catalog. The tags in the example, denoting the properties, are in Spanish, a language that is not necessarily assumed to everybody to know. However, we should remember that semantics in metadata statements can be independent from the spoken language, same way computers do not "speak" English, Spanish or Portuguese either, the same metadata in its behind-the-scenes MARC-encoded format (see Figure 7) can be easily understood by both humans and computers once the meaning (semantics) of each field is defined. The advantages and characteristics of MARC, the Machine Readable, are not new. However, the possibilities of the Linked (Open) Data metadata models and the authorities and vocabularies published as linked data (such as for instance those published by the Library of Congress in Washington - http://id.loc.gov/) are huge for the developments of the Semantic Web and electronic contexts. Sadly, developers are not taken enough advantage of the possibilities of Linked Data and the Semantic Web, not even of the simplest metadata possibilities in HTML. In a different example (see Figure 8), it shown the website of the National Library of Spain, again in Spanish and probably not understood by non-Spanish speakers. However, deeper look at the source code (see Figure 9) can shed some light on the meaning of each object, via a remarkable (and unfortunately less and less common) use of Dublin Core metadata in the source code. The decreasing use of metadata by web developers (and mainly by search engines) is a problem that has also been pointed out elsewhere (e.g., Taylor and Joudrey 2009, Satija and Martínez-Ávila 2014). Dublin core can also be used to describe any kind of resources in a bibliographic record (see Figure 10), and even encoded in RDF XML (See Figure 11). All these preceding examples show that descriptive metadata can describe information resources and provides access points, links, and navigation for them by categorizing, labeling, and structuring important pieces of information about them in a great variety of ways and contexts, especially when aiming at semantic representation in electronic environments. In any case, the representation can be processed/understood as statements

(property/value pairs) about the resources. While the categories and labels differ from one standard to another, for example, AACR2 and MARC, Dublin Core, VRA Core, etc., the meaning of these labels must be stated somewhere. The meaning (semantics) of these vocabulary can be made explicit via authority files and linked data. The metadata content may be encoded in different machine-readable or markup language schemes, for example, MARC, HTML, XML. The encoded metadata may reside in a catalog, database, or database-driven Web site, or it may alternatively be embedded within the resource itself (as in case of an HTML Web page header or an MS Word document). In the case of program objects and applications, these metadata can be also embedded within the objects.

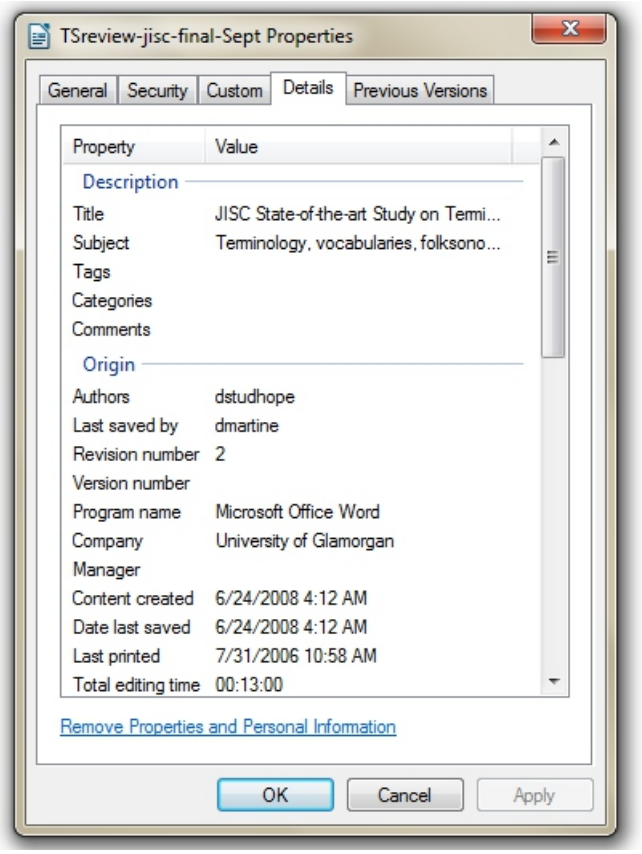


Figure 4: Example of descriptive metadata embedded in the "properties" of a Microsoft Word document

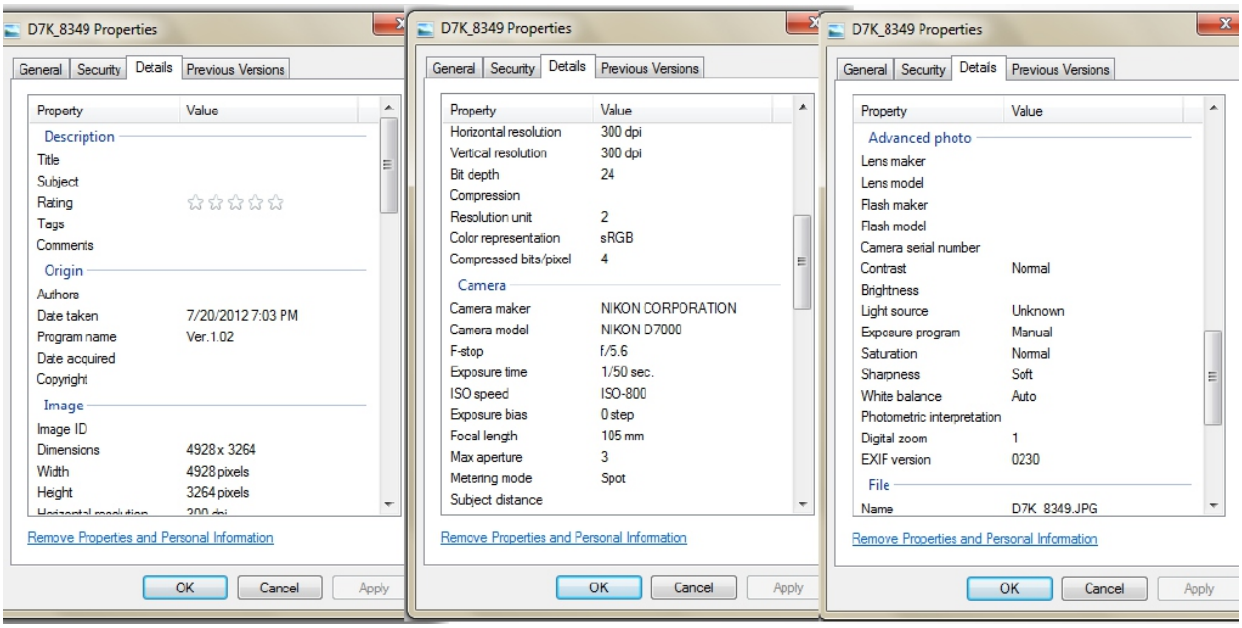


Figure 5: Example of mostly technical metadata automatically generated by a digital camera for a digital photograph

Ver signatura/s | Registro del catálogo

Advances in geographic information systems and remote sensing for fisheries and aquaculture : summary version

Cambiar visualización

Meaden, Geoffrey J.

Cabecera am c0n a
Clave a5388138
Fecha/hora última transc. 20131028
Cód. inform. gen. 130821s2013 it a
ISBN 978-92-5-107391-9
N.º control sis. pro. (OCoLC)864328198
Fuente de catalogación SpMaBN spa SpMaBN rdc
CDU 913(100):004
CDU 639.2:658.011.8
CDU 639.3/.5:658.011.8
CDU 528.88:639.2/.5
Título Advances in geographic information systems and remote sensing for fisheries and aquaculture : summary version / edited by Geoffery J. Meaden and José Aguilar-Manjarrez
Publicación Rome : Food and Agriculture Organization of the United Nation, 2013
Descripción física X, 98 p. : il. col., mapas ; 25 cm + 1 CD-ROM
Tipo de contenido Texto (visual)
Tipo de medio sin mediación
Serie (FAO fisheries and aquaculture technical paper, ISSN 2070-7010 ; 552)
Bibliografía Bibliografía: p. 91-94
Encabez. materia Sistemas de información geográfica
Encabez. materia Tecnología pesquera
Encabez. materia Acuicultura -- Innovaciones tecnológicas
Encabez. materia Teledetección -- Aplicaciones
Autor Meaden, Geoffrey J.
Autor Aguilar-Manjarrez, José
Mantenido por BNALCALA

Figure 6: Example of descriptive metadata as displayed to users of a library catalog (in Spanish)

Ver signatura/s | Registro del catálogo

Advances in geographic information systems and remote sensing for fisheries and aquaculture : summary version

Cambiar visualización

Meaden, Geoffrey J.

```
000       am c0n a
001       a5388138
005       20131028
008       130821s2013 it a | ||||o000 0 eng
020       978-92-5-107391-9
035       (OCoLC)864328198
040       SpMaBN|bspa|cSpMaBN|erdc
080       913(100):004
080       639.2:658.011.8
080       639.3/.5:658.011.8
080       528.88:639.2/.5
245 00    Advances in geographic information systems and remote sensing for fisheries and aquaculture :|bsummary
          version|cedited by Geoffery J. Meaden and José Aguilar-Manjarrez
260       Rome|bFood and Agriculture Organization of the United Nation|c2013
300       X, 98 p.|bil. col., mapas|c25 cm|e1 CD-ROM
336       Texto (visual)|2isbdcontent
337       sin mediación|2isbdmedia
490 0     FAO fisheries and aquaculture technical paper|x2070-7010|v552
504       Bibliografía: p. 91-94
650  7    Sistemas de información geográfica
650  7    Tecnología pesquera
650  7    Acuicultura|xInnovaciones tecnológicas
650  7    Teledetección|xAplicaciones
700 1     Meaden, Geoffrey J.
700 1     Aguilar-Manjarrez, José
956       2
```

Figure 7: Same metadata in its behind-the-scenes MARC-encoded format

Figure 8: an example of a website (in Spanish)

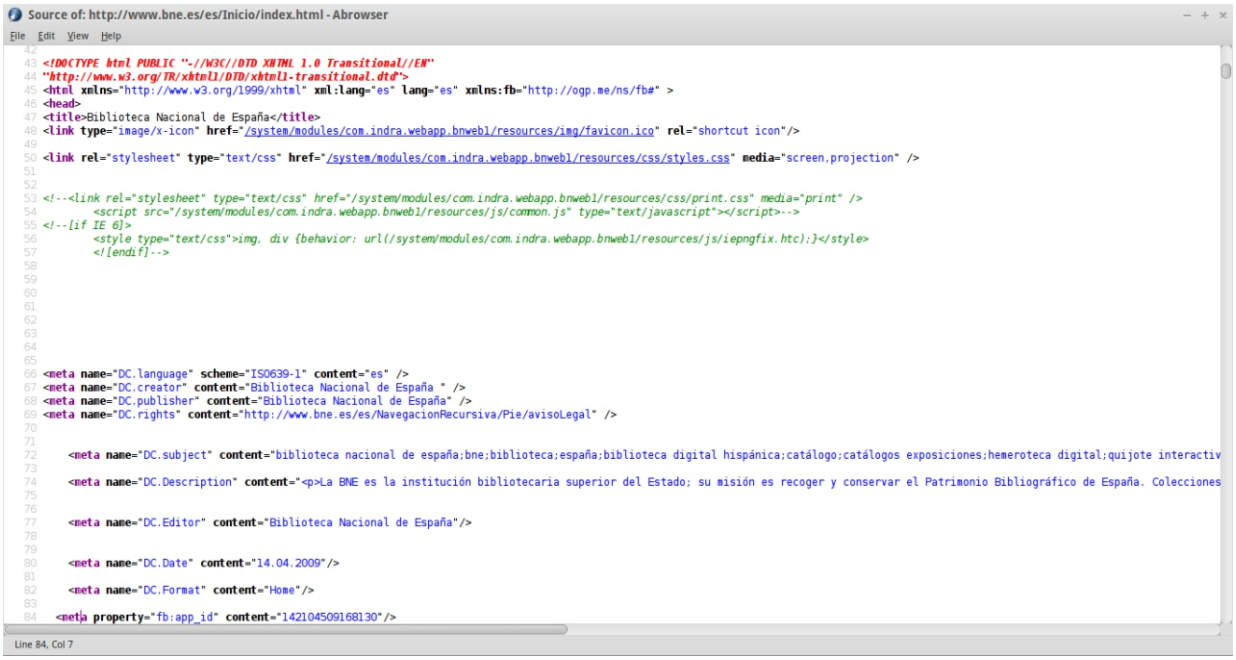


Figure 9: descriptive metadata (Dublin Core) about the Spanish website embedded in the HTML source code

DC.Title	Aerial view of O'Hare International Airport in Chicago, Illinois
DC.Description	O'Hare International Airport; Northward, air view; Chicago
DC.Creator	Mayer, Harold
DC.Subject	Air Transportation
DC.Subject	Airports ; Roads
DC.Subject	Aerial views
DC.Coverage	North and Central America
DC.Coverage	United States
DC.Coverage	Illinois
DC.Coverage	Chicago
DC.Date.Created	1974
DC.Type	Image
DC.Source	Harold Mayer Collection
DC.Source	35 mm
DC.Source	Color slide
DC.Identifier	8d, 32-203
DC.Rights	Donated by Florence Mayer, Harold Mayer's wife
DC.Source	American Geographical Society Library, University of Wisconsin-Milwaukee Libraries
DC.Rights	The Board of Regents of the University of Wisconsin System
DC.Publisher	University of Wisconsin-Milwaukee Libraries
DC.Identifier	me000027

Figure 10: metadata in a simple Dublin Core tagged form

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.0/"
xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
<rdf:Description about="8d, 32-203">
<dc:title>Aerial view of O'Hare International Airport in Chicago, Illinois</dc:title>
<dc:coverage>North and Central America</dc:coverage>
<dc:coverage>United States</dc:coverage>
<dc:coverage>Illinois</dc:coverage>
<dc:coverage>Chicago</dc:coverage>
<dc:creator>Mayer, Harold</dc:creator>
<dc:publisher>University of Wisconsin-Milwaukee Libraries</dc:publisher>
<dc:date>1974 </dc:date>
<dc:description>O'Hare International Airport; Northward, air view; Chicago</dc:description>
<dc:identifier>8d, 32-203</dc:identifier>
<dc:identifier>me000027</dc:identifier>
<dc:language>und</dc:language>
<dc:subject>Air Transportation</dc:subject>
<dc:subject>Airports ; Roads</dc:subject>
<dc:subject>Aerial views</dc:subject>
<dc:rights>Donated by Florence Mayer, Harold Mayer's wife</dc:rights>
<dc:rights>The Board of Regents of the University of Wisconsin System</dc:rights>
</rdf:Description>
</rdf:RDF>
```

Figure 11: same simple Dublin Core metadata encoded in RDF XML

Types of metadata and their functions

Of the previous examples, although all of them share some characteristics such as the formal structure of the semantic statements, it can also be extracted some differences and typologies. It should be noted though that this classifications of metadata is not mutually exclusive. These types (NISO 2004, p.1) are: descriptive metadata, structural metadata and administrative metadata, while administrative metadata can also be divided into technical metadata, rights management (restrictions) metadata and preservation metadata. According to the NISO, descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords. These data elements are used for the indexing, discovery, and identification of information resources not only by machines but primarily bu humans. They allow users to find and identify resources due to the interpretation (background knowledge) of a semantics that can be well defined or not. When used by or

with the assistance of machines they can retrieve resources individually or in sets. Structural metadata, according to the NISO (2004, p.1) indicates how compound objects are put together, for example, how pages are ordered to form chapters. In electronic contexts, this is the information on the internal organization of the digital information resources. These metadata are often used to display and navigate complex digital objects. For example, if a book has each page scanned as a separate image file, the structural metadata relates each file to each other and to the whole allowing navigation through the book, page-by-page, or jumping to specific points. The administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data that sometimes are listed as separate metadata types such as: rights management (restriction) metadata, which documents and deals with copyright status, ownership, restrictions on access and use, etc. for instance including legally-binding licensing agreements or controlling access to authenticated users; preservation metadata, which contains information needed to archive and preserve a resource, it is the metadata needed for the long-term preservation, integrity, and usability of information resources including documentation on technical specifications, provenance, history of revisions (data refreshing, reformatting), etc.; and technical metadata, that is metadata about the technical aspects of information resources, e.g., digital file characteristics, file format, scanning resolution, hardware and software used, compression information, pixel dimensions, etc. It is very common that library and information schools almost only exclusively focus on teaching descriptive metadata, since that has been assumed to be the traditional task of librarians. However, the acknowledging the existence and importance of the other kinds of metadata in electronic environments, the possibilities for library and information professionals are also broaden as well as hopefully the use and understanding of the semantic importance of metadata in electronic contexts, such as for instance the world wide web, might be improved. Another very illustrating typology of metadata, according to their functions, and that can help broaden the field and awareness of library and information professionals, was given by Gilliland (2008, p. 9):

Table 1: Different Types of Metadata and Their Functions

Type	Definition	Examples
Administrative	Metadata used in managing and administering collections and information resources	• Acquisition information • Rights and reproduction tracking • Documentation of legal access requirements • Location information • Selection criteria for digitization
Descriptive	Metadata used to identify and describe collections and	• Cataloging records • Finding aids • Differentiations between versions

	related information resources	• Specialized indexes • Curatorial information • Hyperlinked relationships between resources • Annotations by creators and users
Preservation	Metadata related to the preservation management of collections and information resources	• Documentation of physical condition of resources • Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration • Documentation of any changes occurring during digitization or preservation
Technical	Metadata related to how a system functions or metadata behaves	• Hardware and software documentation • Technical digitization information, e.g., formats, compression ratios, scaling routines • Tracking of system response times • Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of collections and information resources	• Circulation records • Physical and digital exhibition records • Use and user tracking • Content reuse and multiversioning information • Search logs • Rights metadata

On the other hand, the NISO (2004, pp. 1-2) lists five basic functions of metadata: resource discovery, and here it is said that metadata serves the same functions in resource discovery as good cataloging does by allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information; organizing electronic resources, e.g., aggregating websites according to audience or topic, in other words, building these pages dynamically from the metadata stored in databases instead of the HTML code; interoperability, and I will talked more about this aspect in the next section; digital identification, since most metadata schemes include elements such as standard numbers to uniquely identify the work or object to which the metadata refers, such as the URL (Uniform Resource Locator), PURL (Persistent URL), and DOI (Digital Object Identifier); and archiving and preservation, provided the growing concern that digital resources will not survive in usable form into the future, because digital information is fragile, it can be corrupted or altered, intentionally or unintentionally and it may become unusable as storage media and hardware and software technologies change. Regarding this, metadata are key to ensuring that resources will survive and continue to be accessible into the future. In this vein, archiving and preservation require special elements to track the lineage of a digital object (where it came from and how it has changed over time) to detail its physical characteristics, and to document its behavior in order to emulate it on future technologies.

Metadata as surrogates and the need of standardization

In the tradition of Library and Information Science, metadata have commonly used as surrogates, and usually in a whole piece that rarely has been studied or understood according to the logical units. In this

vein, a set of metadata statements (property/value pairs) about the same resource might be bundled into a metadata record (in a database or other context) so a record would be comprised of a collection of statements. Each metadata record stands as a surrogate in place of the actual resource it represents to users. Metadata statements and surrogate records are a form of information representation. They represent the resource by means of selected properties, but the metadata records themselves are not the actual resource itself. These metadata records can help users find, identify, select, and obtain information resources, whether tangible or digital. However, depending on the semantics and/or implicit knowledge of the subjects these surrogates might be valid only for humans with great amount of background knowledge, not for machines. On the other hand, A surrogate is regarded to be a representation of an item and not the item itself. Surrogates have three functional aspects: description of the item represented; as access points for finding the item (allowing the item to be filed in multiple places – by author name, by subject, under multiple subjects); and as encoding mechanism to serve as a carrier for the description and access points. Summarizing, surrogate metadata records can serve for two big options: 1) Information/knowledge representation, to represent resources to the user, serve as a surrogate for the resource itself, provide descriptive information, and help user identify, evaluate and select; and 2) Information/knowledge retrieval, providing means for searching, browsing, navigation, for known item searches and exploratory searches, to retrieve sets of results, not just individual items grouped according to one or more common characteristics and help user find, collocate, navigate, and obtain the information. To help understand the characteristics of these metadata records, both at structural and semantic levels, surrogates are (or at least must be) standardized.

According to the American National Standards Institute (ANSI, s.d.) "A standard is a document, established by consensus that provides rules, guidelines or characteristics for activities or their results. (As defined in ISO/IEC Guide 2:2004)." Surrogates must be standardized for the following reasons: to make surrogates complete and predictable for finding and advisory functions, to make surrogates compatible with other surrogates and therefore suitable for gathering in a database and to allow exchange of data. In a broader sense, standards play a key role in the exchange of information by allowing finding, making the process predictable (so that we know we can search by author or title or other standard aspects); gathering, adding consistency to bring items with common aspects together (by always naming them in the same way); and advisory, also allowing comparability between different surrogates as to topics covered, date, length, etc. However, it is also true that different schemes may serve distinct needs and audiences here is where interoperability comes into play. In rough and general words, interoperability is the ability of multiple systems, using different hardware and software platforms, data structures, and interfaces, to exchange and share data. However, here the question is with all the metadata standards, initiatives, extensions, and profiles, how can interoperability be ensured?

The shortest answer is that the interoperability and exchange of metadata is further facilitated by metadata crosswalks. A crosswalk is a mapping of the elements, semantics, and syntax from one metadata scheme to those of another. A crosswalk allows metadata created by one community to be used by another group that employs a different metadata standard. The degree to which these crosswalks are successful at the individual record level depends on the similarity of the two schemes, the granularity of the elements in the target scheme compared to that of the source (the mapping of schemes with fewer elements -less granularity- to those with more elements -more granularity- is problematic), and the compatibility of the content rules used to fill the elements of each scheme. Table 2 shows an example of metadata crosswalk mapping provided by the NISO (2004, p. 12), providing equivalences between diverse properties (fields) and standards:

Table 2 Example of Metadata Crosswalk Mapping

	Dublin Core	EAD	MARC 21
Title Element	Title	<titleproper>	245 00$a (Title Statement/Title proper)
Author Element	Creator	<author>	700 1#$a (Added Entry--Personal Name) (with $e=author) 720$a (Added Entry–Uncontrolled Name/Name) (with $e=author)
Date Created Element	Date.Created	<unitdate>	260 ##$c (Date of publication, distribution, etc.)

In practice, there are many standards for many contexts that have been usually adding a bit of confusion to the mix. Examples of general schemes are the DCMES, Dublin Core Metadata Element Set, and MODS, Metadata Object Description Schema. Examples of specialized/domain-specific schemes are VRA, Visual Resources Association Core Categories, EAD, Encoded Archival Description, LOM: Learning Object Metadata, TEI, Text Encoding Initiative (TEI Header), and CSDGM, Content Standard for Digital Geospatial Metadata. Examples of metadata encoding schemes are SGML, Standard Generalized Markup Language, XML, eXtensible Markup Language, MARC, Machine-Readable Cataloging and any other database software encoding, whether proprietary or free software (open source) such as Microsoft Access, DSpace, specific content management systems, etc. The following table 3 shows a typology of data standards including examples of diverse metadata standards for different information communities and listing four types of standards: structure standards, value standards, content standards and format/technical interchange standards (Gilliland 2008, p.3, based on Boughida 2005).

Table 3 Typology of Data Standards

Type	Examples
Data structure standards (metadata element sets, schemas). These are "categories" or "containers" of data that make up a record or other information object.	The set of MARC (Machine-Readable Cataloging format) fields, Encoded Archival Description (EAD), Dublin Core Metadata Element Set (DCMES), Categories for the Description of Works of Art (CDWA), VRA Core Categories
Data value standards (controlled vocabularies, thesauri, controlled lists). These are the terms, names, and other values that are used to populate data structure standards or metadata element sets.	Library of Congress Subject Headings (LCSH), Library of Congress Name Authority File (LCNAF), LC Thesaurus for Graphic Materials (TGM), Medical Subject Headings (MeSH), Art & Architecture Thesaurus (AAT), Union List of Artist Names (ULAN), Getty Thesaurus of Geographic Names (TGN), ICONCLASS
Data content standards (cataloging rules and codes). These are guidelines for the format and syntax of the data values that are used to populate metadata elements.	Anglo-American Cataloguing Rules (AACR), Resource Description and Access (RDA), International Standard Bibliographic Description (ISBD), Cataloging Cultural Objects (CCO), Describing Archives: A Content Standard (DACS)
Data format/technical interchange standards (metadata standards expressed in machine-readable form). This type of standard is often a manifestation of a particular data structure standard (type 1 above), encoded or marked up for machine processing.	MARC21, MARCXML, EAD XML DTD, METS, MODS, CDWA Lite XML schema, Simple Dublin Core XML schema, Qualified Dublin Core XML schema, VRA Core 4.0 XML schema

The multiplicity of standards and continuous tension between specific standards for specific contexts that do no serve well for other contexts and general standards that interoperate well but do not serve well for specific contexts is one of the biggest problems for information organizers and developers. In this vein, standards are sometimes seen as useless apparatus that sometimes slow down the development and adaptability of devices. In the GNU/Linux community, for instance, there was an unfortunate quote of Linus Torvalds, the person who created the kernel Linux for the operating system GNU/Linux, charging against standards in a no very polite way: "THE USER DOESN'T CARE. Pushing the blame around doesn't help anybody. The only thing that helps is Fedora being helpful, not being obstinate. Also, the

the fact is, that from a Q&A standpoint, a memcpy() that 'just does the right thing' is simply _better_. Quoting standards is just stupid, when there's two simple choices: 'it works' or 'it doesn't work because bugs happen'. Standards are paper. I use paper to wipe my butt every day. That's how much that paper is worth. Reality is what matters. When glibc changed memcpy, it created problems. Saying 'not my problem' is irresponsible when it hurts users. And pointing fingers at Adobe and blaming them for creating bad software is _doubly_ irresponsible if you are then not willing to set a higher standard for your own project. And 'not my problem' is not a higher standard. So please just fix it." (Torvalds, 2010). This argument was published on the Red Hat operating system forum on bugs, while discussing a "strange sound on mp3 flash website". This example shows that the practicality and, perhaps disrespect to standards, happens at the developer lever even between top programmers, sometimes caused by commercial agendas (such as in the case of Microsoft and their attempt to monopolize the web browsing with Internet Explorer in the 1990s, until the emergence of Netscape and, from its ashes Mozilla, to the completely insult to the users' rights and global usability by Apple and their continuous push for closed products), and sometimes for lack of global infrastructure. It should not be forgotten that in the development (and sabotage) of standards there are many organizations involved: associations such as the IFLA, institutions such as the Library of Congress, standards organizations such as the NISO (Z39), and sadly, commercial sources such as Adobe, Google, and OCLC too. The fact that the incident with Torvalds happened within the Linux community, presupposed to be more ethical and respectful for aspects such as the neutrality of the Web, freedom of users and non-privative interests is even more striking. This aspects might suggest a relative ignorance about standards even among software developers, however, standards and metadata are everywhere in GNU/Linux too. In the following example, I proceed to analyze the use of some metadata within one of the most iconic, sponsorized by the Free Software Foundation, GNU/Linux operating systems: Trisquel.

Use of metadata in Trisquel GNU/Linux applications

According to The Trisquel Project (2004-2014) "Trisquel is a free operating system, based on GNU/Linux. We mean 'free' in the freedom sense. Our users may, with any part of Trisquel, or even all of it: run; study / audit / modify; copy / distribute, distribute modified versions. These freedoms support users' liberty, privacy, and market-freedom." These freedoms and ethical commitments are based Richard Stallman's (ex-MIT, founder of the GNU Project and the Free Software Foundation and, among other things, holder of more than fourteen honorary doctorates and professorships) famous four freedoms (Stallman 2002, p. 43) that have inspired from copyleft to a vast number of ethical initiatives in information. Indeed, as for 2014, Trisquel was one of the eight GNU/Linux distros accepted and listed as 100% free as in freedom distributions (Free Software Foundation 2014) out of the 285 active distros that are currently listed by Distrowatch (Byfield 2014). The ethical advantages and technical possibilities of

the use of free software and 100% free GNU/Linux distributions in relation to information literacy and other aspects of information science have been widely aired in, for instance, (Martínez-Ávila and Prieto-Gordo 2009).

As an example, the organization of Trisquel's applications in the desktop uses categories and metadata. In Trisquel, applications belong to categories that are specified in a standard file called "Desktop Entry." The specification of this type of file is published by freedesktop.org (2014a) and different desktop environments such as GNOME (e.g., Trisquel) and KDE implement it. According to their website "freedesktop.org is not a formal standards organization, though some see a need for one that covers some of the areas we are working on. For Linux operating system standards, look at the Linux Standard Base project. The X.Org Foundation and the IETF are other groups that do formal standards. The Free Standards Group is one group that publishes 'de jure' standards for free software; freedesktop.org is loosely affiliated with the FSG. Unlike a standards organization, freedesktop.org is a 'collaboration zone' where ideas and code are tossed around, and de facto specifications are encouraged. The primary entry to these discussions is the xdg mailing list" (Freedesktop.org 2013). These statements are just reflections of the multiplicity of standard organizations, and potential practical problems, even in such a restricted area as the GNU/Linux and the free software community. This type of files is written by developers or application packagers and are saved with the extension .desktop. When the application is installed with a package management system, this type of file is installed in a system folder used specifically for them: /usr/share/applications. However, a final user can modify the menu or category assigned to the application (i.e., organize the information at user level) using the tool Alacarte, that in the latest versions of Trisquel is available at Menu – System Settings – Main Menu (see Figure 12).

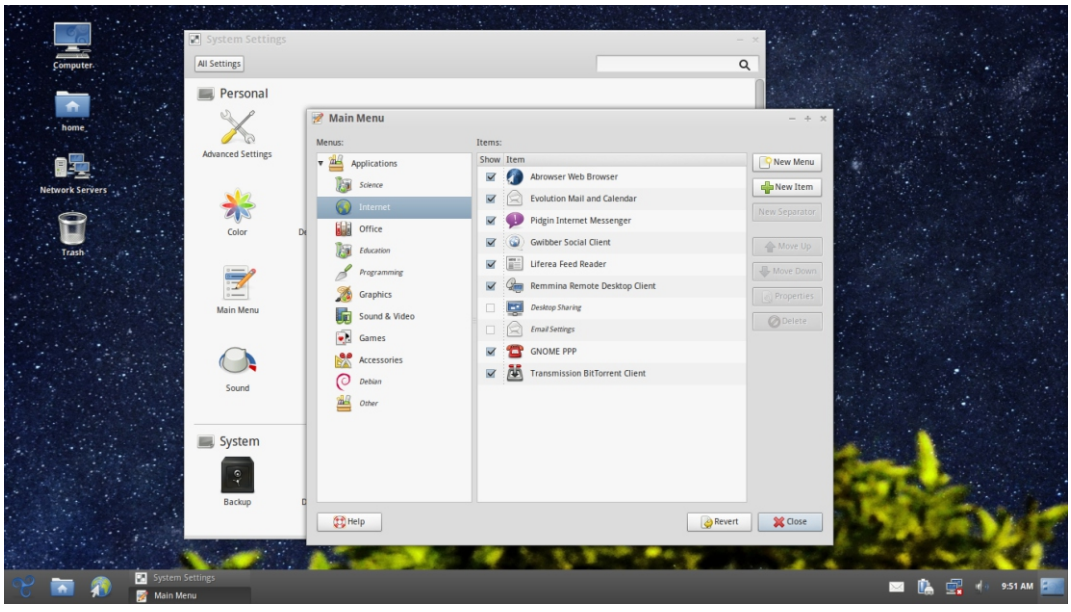


Figure 12 Example of categorization of applications using Alacarte

The properties listed as recognized desktop entry keys by the freedesktop.org specification (Freedesktop.org 2014b) include: Type, Version, Name, GenericName, NoDisplay, Comment, Icon, Hidden, OnlyShowIn, NotShowIn, DbusActivatable, TryExec, Exec, Path, Terminal, Actions, MimeType, Categories, Implements, Keywords, StartupNotify, StartupWMClass, and URL. Out of the previous properties, two of special interest for the purpose of the present study are Categories and Keywords (see Table 4):

Table 4: excerpt of Standard Keys (Freedesktop.org 2014b)

Key	Description	Value Type	REQ?	Type
Categories	Categories in which the entry should be shown in a menu (for possible values see the Desktop Menu Specification - http://www.freedesktop.org/wiki/Specifications/menu-spec/).	string(s)	NO	1
Keywords	A list of strings which may be used in addition to other metadata to describe this entry. This can be useful e.g. to facilitate searching through entries. The values are not meant for display, and should not be redundant with the values of Name or GenericName.	string(s)	NO	1

Sadly, as it can be noted in the previous table, both metadata properties Categories and Keywords are not required. Furthermore, in the following example provided by Freedesktop.org (2014c) both properties are not even included:

```
[Desktop Entry]
Version=1.0
Type=Application
Name=Foo Viewer
Comment=The best viewer for Foo objects available!
TryExec=fooview
Exec=fooview %F
Icon=fooview
MimeType=image/x-foo;
Actions=Gallery;Create;

[Desktop Action Gallery]
Exec=fooview --gallery
Name=Browse Gallery

[Desktop Action Create]
Exec=fooview --create-new
Name=Create a new Foo!
Icon=fooview-new
```

Regarding this, it should be noted that the misuse (omission) of these fields might be caused by the fact that Debian packages (.deb), which is the file type in which applications and other files are packaged in Trisquel, being Debian the ultimate GNU/Linux distro Trisquel is derived from, have a file called "control," similar to the .desktop files, with a field called "section" which works similarly to the "categories" field in ".desktop" files so (see specification of the control file) is specified. This means that although “categories” and “keywords” are not assigned by the developers it is very possible that applications installing packages (such as Add/Remove, Synaptic, Aptitude, etc.) use this file "control" to automatically categorize the packages, while applications such as Alacarte and Main Menu use files ".desktop" that already available in the file system (at /usr/share/applications). However, it should be noted that although the property “categories” is not included in the Freedesktop.org example, in practice it is indeed very common, as it can be checked by opening any ".desktop" file located at "/usr/share/applications" with any text editor, such as for instance gedit (see Figure 13). As for the property “keywords,” sadly, it seems that it is not very used at all.

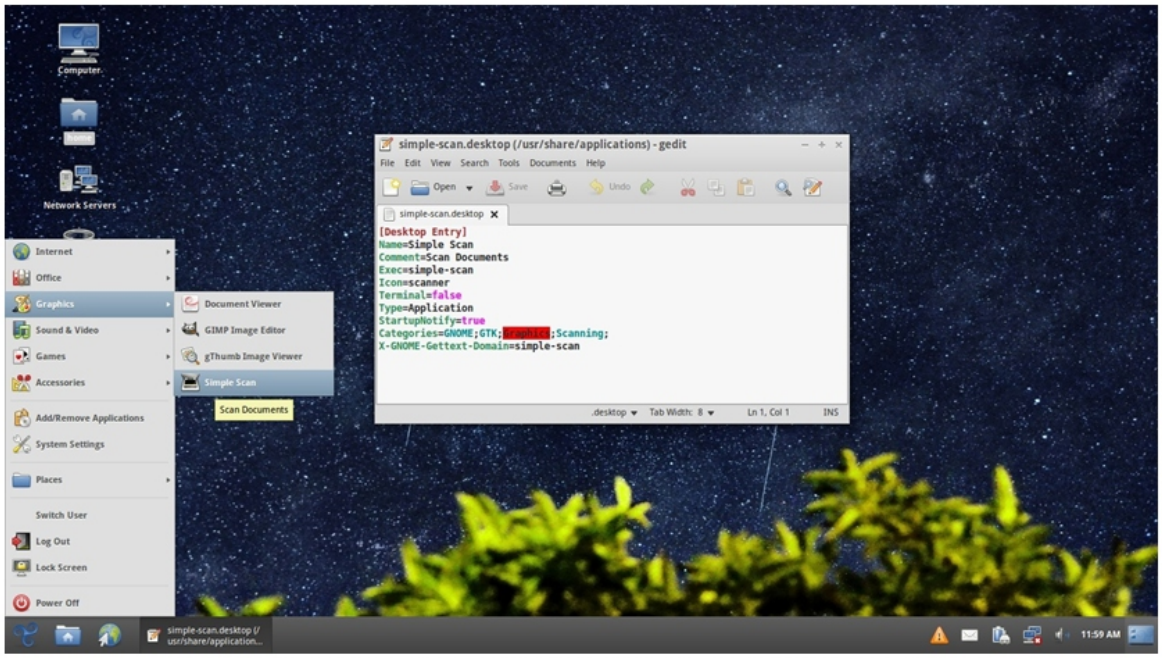


Figure 13: Example of visualization of application categories using gedit

CONCLUSIONS

Despite all the different meanings for the different communities, metadata, as it can understood by the LIS community at its broadest sense, are everywhere. In this vein, the possibilities of providing good semantics and other functions are vital for the good organization of information in electronic environments. Many of today's most promising technologies related to information, especially those

related to the Semantic Web and the description of digital objects, rely on good (standardized) metadata and good descriptions. Although it is commonly possible to do this at the user level, these descriptions are more effective when done by developers and authors. However, several factors such as the multiplicity of schemes and standard organizations, misinformation, and perhaps commercial agendas, are being detrimental to the proper use of metadata in electronic contexts, even among the most ethically committed communities.

REFERENCES

Ansi. S.D. Frequently Asked Questions. Retrieved from http://www.ansi.org/about_ansi/faqs/faqs.aspx?menuid=1

Boughida, K. (2005). CDWA Lite for Cataloging Cultural Objects (CCO): A new XML schema for the cultural heritage community. In Humanities, Computers and Cultural Heritage: Proceedings of the XVI International Conference of the Association for History and Computing: 14–17 (September 2005). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Byfield, B. (2014). Why is the Number of linux distros declining? Retrieved from http://www.datamation.com/open-source/why-is-the-number-of-linux-distros- declining.html

Coyle, K. (2010). Understanding the Semantic Web: Bibliographic Data and Metadata. Library Technology Reports (January 2010). American Library Association.

Freedesktop.org. (2014a). Desktop Entry Specification. Retrieved from http://standards.freedesktop.org/desktop-entry-spec/latest/index.html

Freedesktop.org. (2014b). Recognized desktop entry keys. Retrieved from http://standards.freedesktop.org/desktop-entry-spec/latest/ar01s05.html

Freedesktop.org. (2014c). A. Example Desktop Entry File. Retrieved from http://standards.freedesktop.org/desktop-entry-spec/latest/apa.html

Free Software Foundation. (2014). Free GNU/Linux distributions. Retrieved from http://www.gnu.org/distros/free-distros.html

Gilliland, A. J. (2008). Setting the stage. In Murtha Baca (Ed.), Introduction to Metadata. Online Edition, Version 3.0. Los Angeles: J. Paul Getty Trust. Retrieved from http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting. Pdf.

Martínez-Ávila, D. & Prieto G. O. (2009). ¿Libertad o pragmatismo?: el papel del software libre en la alfabetización tecnológica. In II Conferencia Internacional sobre Brecha Digital e Inclusión Social, 28 al 30 de Octubre de 2009. Madrid: Universidad Carlos III de Madrid, Instituto Universitario Agustín Millares de Documentación y Gestión de la Información.

National Information Standards Organization. (2004). Understanding metadata. Bethesda, MD: National Information Standards Organization.

Satija, M.P. & Martínez-Ávila. D. (2014). Use of classification in organizing and searching the web. Annals of Library and Information Studies, 61 (4), 294-306.

Stallman, R. M. (2002). Free software, free society: Selected essays of Richard M. Stallman. Boston, MA: Free Software Foundation.

Taylor, A. G. & Joudrey. D. N. (2009). The Organization of information, 3rd edn. Westporet: Libraries Unlimited.

Torvalds, L. (2010). Bug 638477 - Strange sound on mp3 flash website - (In reply to comment #128). Retrieved from https://bugzilla.redhat.com/show_bug.cgi?id=638477#c129

Trisquel Project. (2004-2014). Documentation | Trisquel GNU/Linux - Run free!. Retrieved from http://trisquel.info/en/wiki/documentation.

Design and Analysis of a Big Data Technologies and WI [Web Intelligence] and Using Hadoop: A Tool for Solution

Jitender Rai
Associate Professor, Tecnia Institute of Advanced Studies Email id-
jitender12rai@gmail.com,Contact No-9990365276

A B S T R A C T

This research paper focused on new emerging technologies Big Data has captured a lot of interest in industry, with anticipation of better decisions, efficient organizations and many new jobs. Much of the emphasis is on the challenges of the four V's of Big Data: Volume, Variety, Velocity, and Veracity, and technologies that handle volume, Including storage and computational techniques to support analysis (Hadoop, NoSQL, Map Reduce, etc).We analyzed and incorporate a lot of data in social networks and suggest an appropriate data model. We further provided a recommendation system that can be used together with the data model for high reliability. The Web Intelligence is building 'web-intelligence' applications exploiting big data sources arising social media, mobile devices and sensors, using new big-data platforms based on the 'map-reduce' parallel programming paradigm. The interest in WI is growing very fast. We would like to invite everyone, who are interested in the WI related research and development activities, to join the WI community. Your input and participation will determine the future of WI. Today big data is buzz word that pertain to problem as well as solution involving technology for storing and managing the very large data-set of different type in age of internet, intra net and satellite communication. Most of development in field to big data is coming from side of computer and related technology (software & solution) giants like IBM, Microsoft, SAP & Sybase, Teradata, MongoDB, Solr etc. Emergence of cloud computing, No SEQUEL technologies, new software tools and database systems for large and unstructured datasets, data Lake etc we use the hadoop tools to solve the problem of big data.

Keywords—Big Data, Hadoop, Big data SEQUEL, No SEQUEL, Data Lake, information exchange capacity, cyber and internet, cloud computing, shared storage architect, Amazon Cassandra, Hadoop, tree, map analysis and map reduce

INTRODUCTION

Big Data is defined as a large amount of data which require new technologies and architecture to make possible to extract value from it by capturing and analysis process. Big data is emerged because we are living in a society which makes increasing use of data intensive technologies. Such large size of data it becomes very difficult to performed effective analysis using the existing traditional techniques. Since Big Data is a recent upcoming technology in the market which can bring huge benefits to the business organization. The difficulties can be related to the data capture, storage, search, analytics, sharing and visualization. Big data is the next generation of data warehousing and business analytics. It has many

deep roots and many branches. In fact you speak with most data industry veterans Big data has been around for decades for firms that have been handling tons of transactional data over the years –even dating back to the mainframe era. Data base is increasing, data sets are burgeoning, and becoming bigger and bigger in very fast way. As things changes and technology upgraded, more and more data get added day by day in every field of humanity, economics & commerce, life and science & technology [1][2]. New researches in different fields generate new technology and larger automation of even smallest of things related to our daily activities. The appeal of big data is rising because unprecedented amount of information moves through organization systems. Big data due to its various properties like, velocity, volume, variety, variability, value and complexity.

Some important Characteristics Big Data:

a) Data Volume: The Big Data word in Big data itself define the volume the data existing is in petabyte (1015) and is supposed to increase to zetabyte (1021) in nearby future. Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.

b) Data Velocity: This deals with the speed of data coming from the various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flow and aggregated.

c) Data Variety: Data variety is the measure of the richness of the data representation, Text, image Audio, Vedio, etc. data being produced is not of single category as it not only include the traditional data but also the semi structure data from various resources like web pages, Web Log Files, social media sites, e-mails, documents.

d) Data Value: Data value measures to the usefulness of the data in making decisions. Data science is exploratory and useful in getting to know the data, but analytic science encompasses the predictive power of big data.

e) Complexity: Complexity measures the degree of interconnectedness and independence in big data structure such that a small change in one or few elements can yields very large changes or small change.

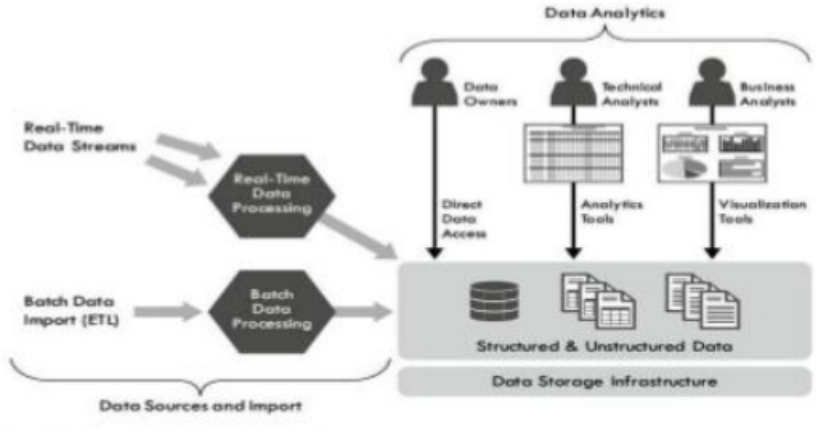


Figure 1.1 Data Analysis architecture of big data

Technologies description of Big Data

A 2011 McKinsey paper suggests suitable technologies include A/B testing, association rule learning, classification, cluster analysis, crowd sourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, predictive modeling, regression, sentiment analysis, signal processing,upervised and unsupervised learning, simulation, time series analysis and visualization. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems [3]. Even though there are many suitable technologies to solve Big Data challenges as indicated in the list above given by the McKinsey paper, this work will explore a handful of the more popular ones, in particular the Amazon Cloud, Hadoop's Map Reduce, and three open-source parsers.[2]

The Elephant in the Room: Hadoop's Parallel World: - Hadoop. We are in a world today where is there is Exabyte's of data being generated every data. Consider the following statistics.
Every minute: Facebook users share nearly 2.5 million pieces of content, Twitter users tweet nearly 400,000 times, Instagram users post nearly 220,000 new photos, YouTube users upload 72 hours of new video content, Apple users download nearly 50,000 apps, and Email users send over 200 million messages, Amazon generates over $80,000 in online sales.

The two critical components of Hadoop are:

1. The Hadoop Distributed File System (HDFS)- HDFS is the storage system for a Hadoop cluster. When data lands in the cluster HDFS breaks it into pieces and distributes those pieces among

the different servers participating in the cluster. HDFS breaks it into pieces and distributes those pieces among servers participating in cluster.

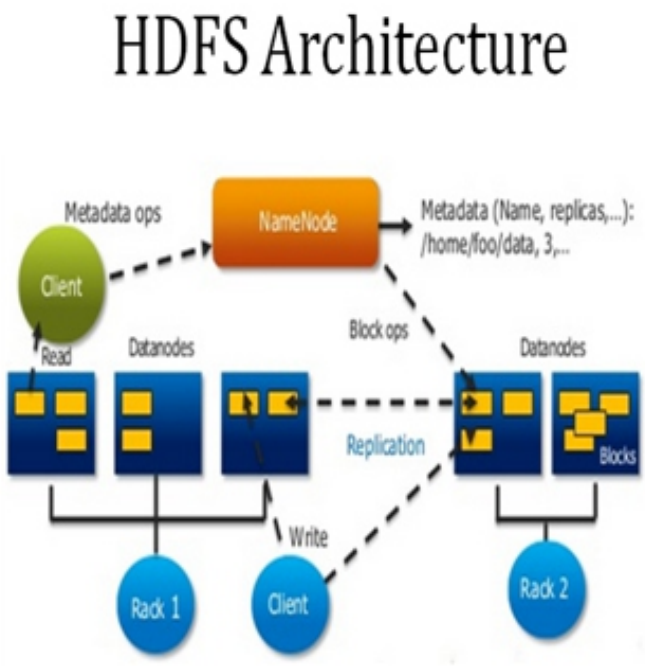


Figure 1.2 HDFS Architecture of big data

2. Map Reduce-Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into comprehensive answer. Map Reduce is the agent that distributes the work and collects the results. Map and Reduce are two functions with shuffle in between which is handled by the system. Both HDFS and Map Reduce are designed to continue to work in the face of system failure. Thus Hadoop provides scalable, reliable and fault-tolerant services for data storage and analysis at very low cost. [3]

Analysis and Survey of Big Data

Big Data analysis tools which are used for efficient and precise data handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. It is interesting to note that for all the tools used, Hadoop over HDFS is the underlying architecture. Oozie and EMR with Flume and Zookeeper are used for handling the volume and veracity of data, which are standard Big Data management tools. The layer with their specified tools forms the bedrock for Big Data management and analysis framework[4].

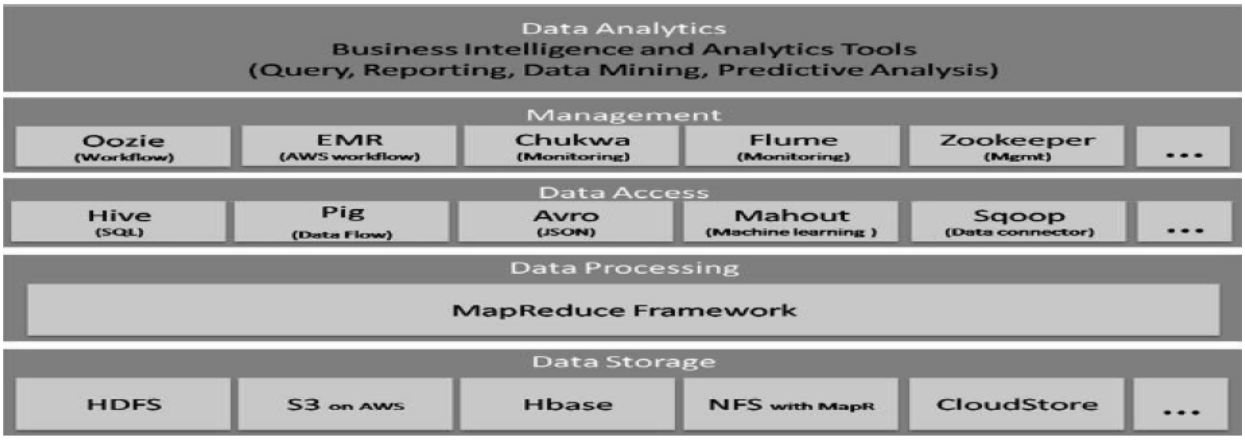


Figure 2.1 Business Intelligence and Analytic Tools of Big data

Roll of Web Intelligence (WI) in Big Data

The WI is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of artificial intelligence (AI), such as knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence, as well as advanced information technology (IT), such as wireless networks etc[4] .The new WI technologies will be determined precisely by human needs in a post-industrial era; namely, information empowerment, knowledge sharing, virtual social communities, service enrichment, and, practical wisdom development. WI techniques and technologies, which cover the following four conceptual levels at least :(1) Internet-level communication, infrastructure, and security protocols. The Web is regarded as a computer-network system. (2) Interface-level multimedia presentation standards. The Web is regarded as an interface for human–Internet interaction [4].

An Intelligent Enterprise Portal Centric Schematic Diagram of WI Technologies:

AI and IT to a totally new domain. On the other hand, the WI technologies are also expected to introduce new problems and challenges to the established disciplines on the new platform of the Web and the Internet. That is, WI is an enhancement or an extension of AI and IT.

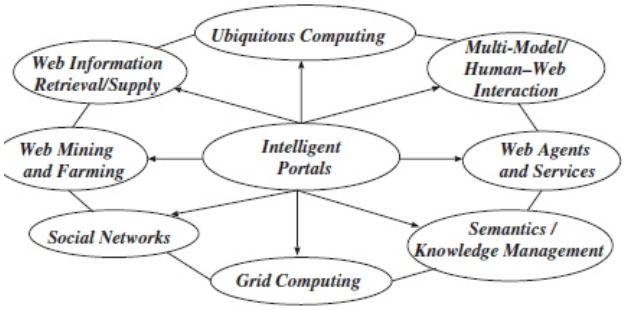


Figure 2.2 An intelligent enterprise portals enteric schematic diagram of WI technologies.

To study advanced WI technologies systematically, and to develop advanced Web-based intelligent enterprise portals and information systems, we provide schematic diagram of WI technologies from a Web-based, intelligent enterprise portals centric perspective[5].

Analysis and Modern Trends

Big data analytics for manufacturing applications are generally based on following points, namely:

a) **Smart connection** – measuring or tapping data y sensors or controller or enterprise manufacturing systems such as ERP, MES, SCM & CMM. A seamless and tether-free technique is required to manage data acquisition procedure and transferring data to central server is required where different protocols like MTConnect etc. are effectively used[6].

b) **Conversion (data to information)** – several tools and methodologies are used at this level.

c) **Cyber level is central information hub.** Specific analytics are used to extract information and these analytic provide machines with self comparison ability with respect to the whole fleet.

d) **Cognition** – full knowledge of monitored system is produced by implementing CPS at this level. Exact info-graphics are properly used at this level to transfer completely the acquired knowledge to users.

e) **Configuration** – it is the feedback from cyber space to physical space and this level act as supervisory control to make machines self-configure and self-adaptive. Hence the level acts as RCS (resilience control system) and it apply the decisions which are corrective and preventive in nature. Such decisions are made at cognition level to the monitored system.

f) The trends of modern technology with respect to big data include: Beginning and development of cloud computing, New software tools and database systems for large but unstructured datasets, Refining analytical tools to process vast quantities of data in near-real time, Monetization of big data sources, Concerns around privacy of data and intellectual property ,Rise of global smart cities.

Shared Storage Architecture-

Shared storage architectures like storage area network (SAN) and network-attached storage (NAS) are typically slow, complex and expensive in big data domain. Big data analytic systems have to perform faster, should have simpler commodity infrastructure and low cost. Hence shared storage system and architectures, since 2011, are not in use in today's world of big data [7]. Real or near-real time information delivery is one of defining characteristics of big data analytics. Big data is difficult to work with using most RDBMS and desktop statistics and visualization packages because it requires 'massively parallel software running to hundreds or thousands of servers'. Some most prominent types

of NoSQL data technologies that are still under way of development include: (a) Storing huge data sets of document in suitable way (b) Emergence and development of wide column store in order to access structured and semi structured data rapidly.(c)Rapid access to unstructured data by key value store (d)Advance search engine suitable for full text indexing for documents (e)More effective graph database for storing graph type data like social networks[8].

Efforts in Big Data Networking

MongoDB (from 'Humongous') is an open-source document database, and leading NoSQL database. It is leading technology in 'document store'. It uses JSON- Style documents with dynamic schemas providing power and simplicity. Apache Cassandra is database technology which is mostly preferred for its scalability and high availability without compromising performance. It is often referred to as a data Cloud analytics is a data driven frontier and it leverages the use of cloud and related computing, communications and data management and visualization technologies to perform sophisticated multivariate analysis on massive centralized, distributed or federated data sets to better understand and help solve problems. Booz Allen is one of global leader and innovator in bringing cloud analytics to commercial clients [10].

Big Data: Hadoop A Tool for Solution

Now we can see that Hadoop-based MBD processing system (HMBDPS) has three components that each covers a particular application, as shown in Fig.3.1.

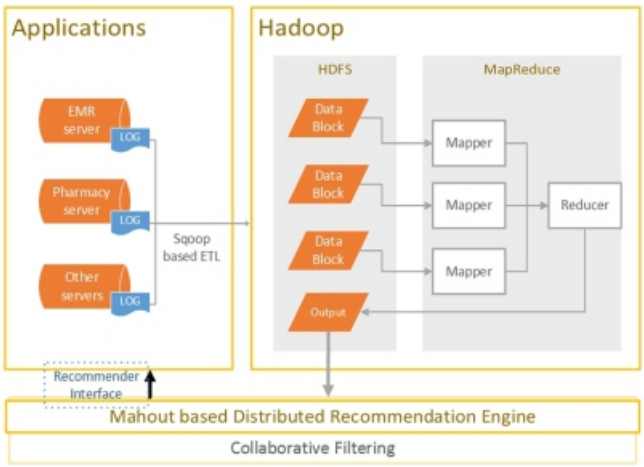


Figure 3.1 Hadoop Architecture

The components are a Sqoop-based Extract-Transform-Load (ETL) module; a Hadoop cluster for data. Storage, management and parallel processing. After Combined with the applications that create a user behavior log and a recommender interface server, all of the components form a closed loop system. We can have look on the functions as follows let's have a look on Data development of healthcare

information technologies, clinical data are typically stored in relational databases. We all know that because relational databases are conventional information technology (IT) architectures, they cannot face the challenges and fix the problems caused by big data. On contrary, Hadoop is a reliable platform for processing big data. Current methods of transferring data are inefficient[12]. In our system, we developed Sqoop-based ETL (SETL) module, which is designed to offer effective and efficient MBD transferring services to optimize our usage of MBD, as shown in Fig. 3.2. Sqoop is a tool for efficiently transferring bulk data between Hadoopand structured datasets, such as a relational database management system (RDBMS). To transfer data, the SET module first identifies a connector to the source data through Java Data Base Connectivity (JDBC), looks up the metadata of the source data, and then transforms Structured Query Language (SQL)-type data to Java-type Sqoop records as an input format for Map Reduce jobs. Finally, appropriate numbers of map tasks and reduce tasks are launched to write the records to the HDFS, as shown in Fig.3.3.

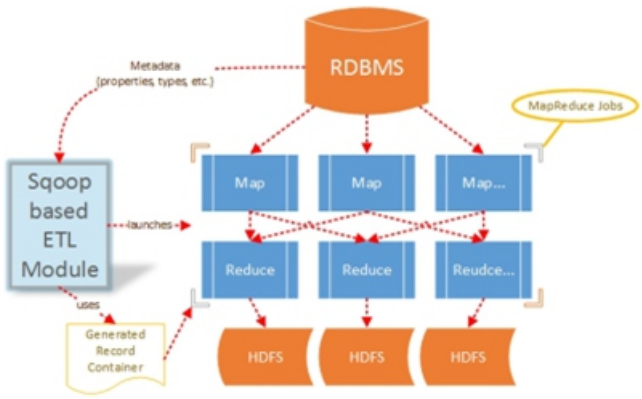


Figure 3.2 HDFS Architecture of Big data

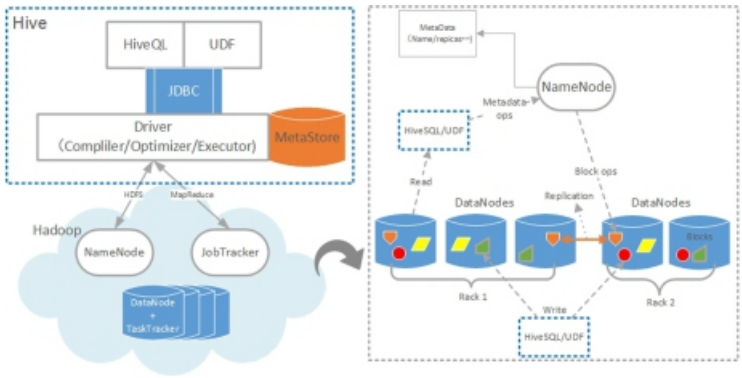


Figure 3.3 Hadoop: Core Component of Engine Architecture of Big Data

CONCLUSION

The conclusion of the paper describes the use of big data and how to related the other emerging technologies like web intelligence and cloud computing so define the major roll of functionality of big data in different areas and also design and development of a Hadoop-based MBD processing system of

big data that can be applied for secondary uses of MBD. This system solves problems of big data using MBD collection, storage and analysis have also seen how Hadoop work. Compared with non-distributed systems, Hadoop demonstrate that this system and its related distributed algorithms are capable of handling MBD more efficiently. Since Hadoop is a distributed system it works more efficient as it has master and slaves parts. Focused on the study of WI is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of artificial intelligence (AI),1 such as knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence, as well as advanced information technology (IT), such as wireless networks; ubiquitous devices; social networks There is increasing number of technologies being in use for cloud –data processing. For example, some cloud based data warehouse and technologies include etc.

REFERENCES

Bala, Nayar Kiran. (2013). Reliability of Cloud Computing Environment during Disasters. Indian Research in Progress. Retrieved from http://shodh.inflibnet.ac.in:8080/jspui/handle/123456789/1625

Booz Allen Hamilton. (2015). Cloud Analytics. Retrieved from http://www.boozallen.com/consulting/technology/cloud-computing/cloud-analytics

Borthakur, Dhruba. (2007). The hadoop distributed file system: Architecture and design. Apache Hadoop Documentation. Retrieved from https://hadoop.apache.org/docs/r0.18.0/hdfs_design.pdf

Chang, Fay, et al. (2008). Bigtable: A distributed storage system for structured data. In ACM Transactions on Computer Systems (TOCS). Retrieved from http://static.googleusercontent.com/media/research.google.com/en/us/archive/bigtable-osdi06.pdf.

Dan Sommer, Rita L. Sallam & James Richardson. (2011). Emerging technology analysis: Visualization-based data discovery tools.

Garcia, Ted & Wang, Taehyung ("George"). (2013). Analysis of Big Data Technologies and Methods. In IEEE Seventh International Conference on Semantic Computing California State University.

Garlasu, Dan, et al, (2013). A big data implementation based on Grid computing. In Roedunet International Conference (RoEduNet), IEEE, 11th, Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6511732

Liu, J. (2003). New challenges in the world wide wisdom web (W4) research. In Zhong, N. et al. (eds.), Foundations of Intelligent Systems, LNAI 2871, pp. 1–6.

Liu, J. (2003). Web intelligence (WI): what makes wisdom web? In Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), pp. 1596–1601.

Liu, J., Zhong, N., Yao, Y & Zbigniew, W. Ras. (2003). The wisdom web: new challenges for web intelligence (WI). Journal of Intelligent Information Systems, 20:1, 5–9.

Manikandan, S.G. & Ravi, S. (2014). Big Data Analysis using Apache Hadoop. In International Conference on IT Convergence and Security (ICITCS), 2014. IEEE, pp.1-4.

Minelli, Michael, Chambers, Michele & Ambiga Dhiraj. (2013). Big data, big analytics : emerging business intelligence and analytic trends for today's businesses. Hoboken, New Jersey : John Wiley & Sons, Inc.

Qubole. (2015). Big Data Cloud Database & Computing. Retrieved from http://www.qubole.com/big-data-cloud-database-computing/

Vael, Marc. (2013). How to manage big data and reap the benefits. ComputerWeekly.com. Retrieved from http://www.computerweekly.com/opinion/How-to-manage-big-data-and-reap-the-benefits

Zhong, N., Liu, J. & Yao, Y. (2002). In search of the wisdom web. IEEE Computer, 35(11): 27–31.

Users' Information Search Strategies in Computerised Libraries of Degree College Sivasagar District, Assam

Hima Devi[1], Jnyandeep Saikia[2]
[1]Research Scholar,
Assam down town University. Assam, India. Email:himadeviblbc@gmail.com
Mobile:9435646831
[2] Librarian,DKDCollege,
Dergaon,Assam. Email:jnyandeeps@rediffmail.com, Mobile.9435771487

A B S T R A C T

The study is focused on finding of the users' information search strategies in computerised degree college libraries of Sivasagar District, Assam.A survey has been conducted for collecting data regarding information search strategies using a well-structuredquestionnaire, which was randomly administered personally among theuser communitycomprising student, teacher and staff of the colleges. The sample size was 170 (Students-100,Teachers -50 and Staff- 20). Based on the feedback of the respondents the data has been analysed and conclusions were made. The result of the study has revealed that 61.75% user visit library daily, 75.88% users use OPAC. 74.71% users awareof the resources, services and facilities of their libraries fully and 22.35% having limited awareness about those.

Keywords—ICT, Academic library, User opinion, Information use pattern.

INTRODUCTION

The present day society needs accurate, authentic, exhaustive and up-to-date information for every sphere. It is the key factor for planning and decision making process. It is also treated s the key component for growth and development of a nation, indication of living standard and as the life blood of a society and so on . The higher education, being the vital instrument of a country, plays an important role in the growth and prosperity of a nation. The colleges being most important higher educational institutions after university of society are set up for the preservation, creation and dissemination of knowledge and information. The academic libraries are built to support teaching, learning, research and other educational functions to their parent institutions. Without a sound library, a higher educational institution cannot achieve the desired goal.

Library and information centers are entirely service oriented organization and completely opposite to other profit making organisations. As such, the success or failure of a library services is judged by the customer's satisfaction. So, the services must be in accordance to the users need /demand or satisfaction.

Any new addition or subtraction of service must depend upon the user's requirement or preferences .This is attained through understanding the users diversified requirements through conducting continuous proper user study.

The user community of a library and their information requirements are heterogeneous in nature. The users come to the library for aspecific purpose and are comparable by one or more criteria, they are having personal characteristics. To understand the user community, we have to study their individual characteristics properly . We know that, a user as an individual may also differ in terms of age, gender, experience; attitude, beliefs, habits, values, goals, interest, motivation, communication, education, specialisation, expertise , capabilities , abilities , cultural background etc.

As the user category are varied, their information seeking behaviourare also a very complex patternand diversified in nature. We must provide quality services to the needy users by understanding the user community and their requirements by taking user survey and analysing their feedbacks time to time.

OBJECTIVES

The study focuses on the users' information search strategies in computerised Degree College Libraries of Sivasagar District, Assam. Some specific objectives are :

- To find out the information use pattern of the users
- To assess the awareness of the user about the resources and facilities of the libraries
- To find out the purpose for which they seek information
- To find the status of computer applications in college libraries
- To ascertain the users responses in computerised college libraries
- To know the attitude of the user community towards computer application in library services

METHODOLOGY

To conduct the study, survey or descriptive method has been adopted and questionnaire along with observation have been chosen as appropriate tools for collecting data.A structured questionnaire was provided to the user categories comprising students, teachers and staff of the respective colleges to collect data . The questionnaires were distributed to the library users (10 for students, 5 for teachers and 2 for staff in each college) present in their reading room, computer labs, UGC resource centre of the libraries and collected their feedback . The sample size was 170 (Students-100, Teachers -50 and Staff-20).The gender wise response rate is 59% male and 41% female.

LIMITATION OF THE STUDY

The study restricts itself to the libraries of the provincialised general degree colleges of Sivasagar District which are affiliated to Dibrugarh University under 2 (F) & 12 (B) of UGC Act.

BACKGROUND OF THE LIBRARIES

There are 14 provincialised degree colleges in the District under Dibrugarh University having 2 (f) & 12 (B) of UGCAct. Out of them, data are collected from 10 colleges (71.43%) for the study.The colleges are Amguri College, Gargaon College, JHNS College , Moran College, Nazira College, Sibsagar College, Sibsagar Girls College, Sonari College, SMD College and SPP College. Among the colleges, the Sibsagar College is the oldest college which was established in the eve of independence of our country.Among the surveyed colleges,100% colleges having Arts stream followed by 60 %Science and 40% commerce stream.90% colleges provide co-education and Post Graduate Courses.All the colleges have been accredited by NAAC. Out of them 80% colleges got B+and B Grade.Most of the colleges practice open access (90%). It is observed that all the colleges automated their libraries using SOUL library software.

Data analysis and findings: After receiving the responses from the target group, data are tabulated and analysed properly to understand the users' information search strategies .Their interpretations have been provided in the successive sub heads.

Information search strategies of the users:

Academic librariesare established by their parent organizations to serve their bonafide users. The user satisfaction is the ultimate goal of the libraries. The success or failure of any library and any addition or subtraction of services or facility completely depend upon users requirements. As such, we have to study the user community, their requirements and take their feedbacks regularly and to provide timely response to them. The feedback of the present study are described below .

Frequency of library Visit :Table 1 reveals that majority of the users 105(61.76%)visit library daily, followed by weekly 46 (27.07%) and sometimes 16(9.41%).

Table1:Nature of Library Visit

User Category	Nature of Library Visit			
	Daily	Weekly	Sometimes	Rarely
Student	72	24	4	
Teacher	31	16	3	
Staff	2	6	9	3
Total	105	46	16	3

Purpose of library visit: Table 2 depicts clearly the purpose of library visit of the users. It reveals that most of the users 125 (73.53%)purpose of library visit is to read booksfollowed by 59.41% to collect text books, 55.29% to brows Internet, 47.06% to consult periodicals, 45.29% to read newspapers and 43.53% to consult reference books.

Table 2 : Purpose of library visit

Purpose	No. of Respondents	Percentage	Rank
To read books	125	73.53%	1
To collect Text Books	101	59.41%	2
To brows Internet	94	55.29%	3
To consult Periodicals	80	47.06%	4
To read Newspapers	77	45.29%	5
To consult Reference Books	74	43.53%	6
To know the latest arrivals	53	31.18%	7
To consult non book resources	33	19.41%	8
Other Purposes	16	9.41%	9

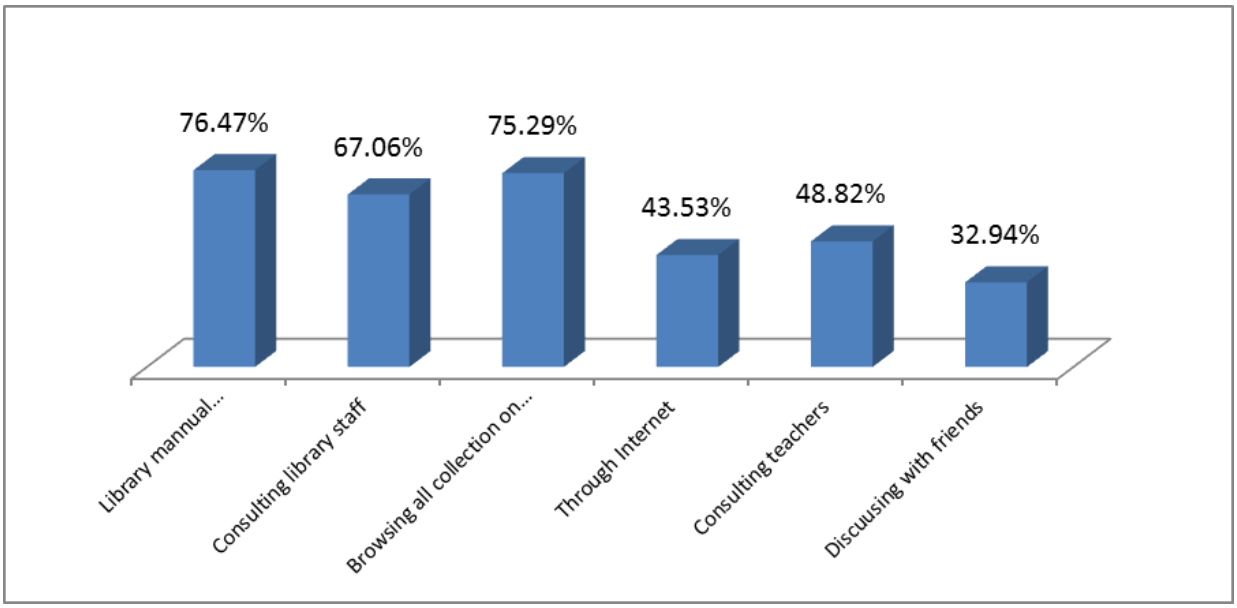
Purpose of seeking Information :The purpose of seeking information of the user community is heterogeneous in nature. The table3 indicates that majority of the users 72.35 % purpose of seeking information is for career enhancement followed by 68.24% for academic assignment, 49.41% to keep up-to-date on subject of interest and 42.35% for learning to prepare for competition.

Table 3: Purpose of seeking information

Purpose	No. of Respondents	Percentage	Rank
For academic assignment	116	68.24%	2
For learning to prepare for competition	72	42.35%	4
To keep uptodate on subject of interest	84	49.41%	3
For career enhancement	123	72.35%	1
Other	41	24.12%	5

Sources to locate Information:There are various sources to locate information. Figure 1 indicates that most of the respondent 130 (76.47%) locate information through library manual catalogue/OPAC. This is followed by 75.29% through browsing all collection of library stacks, 67.06% consult the library staff, 48.82% consult their teachers and 43.53% users find their required information through browsing Internet.

Figure1: Sources to locate information



Use of OPAC by the Users :Table 4 reveals that most of the users 129 (75.88%) use OPAC for searching their requirements. Out of them , the user category teachers use OPAC maximum 82% followed by the students 79% and staff 45%.

Table 4. Use of OPAC by the Users

Users' Category	Use of OPAC by the Users'	
	Yes	No
Student	79	21
Teacher	41	9
Staff	9	11
Total	129	41

Users awareness about the Resources, Services and Facilities of the Libraries: Users awareness about the resources, services and facilities of the libraries is out most important for the best practices of the library. The table 5 shows that most of the users fully Aware 127(74.71%)of the resources, services and facilities of their libraries. It is also observed that some of the users 38(22.35 %) awareness is limited and a few users 2.94% not at all aware those.

Table 5: Users awareness about the resources, services and facilities of the libraries

Users Category	Fully Aware	Awareness is Limited	Not at all Aware
Student	79	20	1
Teacher	46	4	
Staff	2	14	4
Total	127	38	5

Users' demand for Library Home Page /Website for Information Dissemination: The table 6 indicates that there is a strong users demand (85.29%) for library Home Page /Website for dissemination of information. Out of them the students and teachers demand for it is maximum in comparison to the staff user category. Most of the students (95%) and teachers(82%) demand the library Home Page /Website.

Table 6: Users' demand for library Home Page /Website for Information Dissemination

Resources	No. of Respondents	Percentage	Rank
Books	148	87.06%	1
Journals	98	57.65%	3
Newspaper and Magazines	126	74.12%	2
E- Resources	75	44.12%	4

Mostly used Library Resources:Table 7 reveals that majority of the respondents 87.06% use books from their library followed by the newspaper and magazines 74.12%, journals 57.65% and E- resources 44.12% than other library resources.

Table 7 : Mostly used Library Resources

Resources	No. of Respondents	Percentage	Rank
Books	148	87.06%	1
Journals	98	57.65%	3
Newspaper and Magazines	126	74.12%	2
E- Resources	75	44.12%	4

Quality of the Resources of the Libraries:Table 8 shows that majority of the users84 (49.41%) find the resources of the libraries good followed by very good (21.18%), average (18.24%), excellent (10%).

Table 8 : Quality of the Resources of the Libraries

Quality of the Resources	No. of Respondents	Percentage	Rank
Excellent	17	10%	4
Very Good	36	21.18%	2
Good	84	49.41%	1
Average	31	18.24%	3
Poor	2	1.18%	5

Major Findings

- The study reveals that majority of the users 105(61.76%) visit library daily.
- Most of the users 125 (73.53%) purpose of library visit is to read books followed by 59.41% to collect text books, 55.29% to brows Internet.

- The study finds that majority of the users 72.35 % purpose of seeking information is for career enhancement followed by 68.24% for academic assignment, 49.41% to keep up-to-date on subject of interest, 42.35% for learning to prepare for competition.
- Majority of the respondent 130 (76.47%) locate information through library manual catalogue/OPAC. This is followed by 75.29% through browsing all collection on library stacks, 67.06% consult the library staff.
- It is found that most of the users 129 (75.88%) use OPAC for searching their requirements. Among the user category, teachers use OPAC maximum 82% followed by the students 79% and staff 45%.
- Majority of the respondent demand 85.29%for library Home Page /Website for dissemination of information. Most of the students (95%) and teachers(82%) demand the library Home Page /Website.
- The study shows that majority of the respondents 87.06% use books from their library followed by the newspaper and magazines 74.12%, journals 57.65% and E- resources 44.12% than other library resources

CONCLUSION AND RECOMMENDATIONS

The study reveals that majority of the users visit their library daily and use OPAC for searching their requirements. The student community visit the library daily and use maximum the library resources services and facilities . The use of e- resources by the user community is less than other resources. The colleges which practice close access must practice open access for best use of the libraries and to attract the users and make the potential user to actual reader. There must be practiced regular user education programme in the colleges to cater reading habit amongst the user community.

REFERENCES

Abifarin, A (1994). Information seeking behaviour of agricultural students in selected Nigerian universities. Journal of Library And Information Science, 19, (2).

Abraham, J (2012). Information use pattern of enterprenuers of small scale industries (SSI) : A study of Dindigul District. In Dynamics of librarianship in the knowledge society. Vol.2 p.542.

Adika,G.(2003).Internet use among faculty members of universities in Ghana. Library Review, 52(3),29-37.Retrived from http://www.emeraldinsight.com/journals.htm?articleid=859714

Ansode,S.Y.(2012).Assessing ICT literacy skills of research students of University of Pune. AsiaPecific Journal of Library and Information Science,2(1),1-12.Retrieved from http://apilis.msu.ac.th/ojs/index.php/APJLIS/article/view/30/pdf1

Bhattacharjee, S .(2008). Information seeking behavior of degree students of general colleges of Barak Valley, An unpublished PhD thesis, Gauhati University.

Biswas,S.C., and Choudhury(Deb Roy),T.(2011).Role of Information needs in developing collection of college libraries with special reference to Barak Valley : A case study. In national seminar on collection development. Silchar : Assam university, Feb.10-11.

Choukhande,V.G.(2008). Information needs and information seeking behaviour. Amravati :Shivneri Publisher.p113.

Ellis, D. and Haugan, M.(1997). Modeling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation,53(4), 384-403.

Gautam, A. K.. (1992). Modern Information Technology and Reader"s Services in

University Libraries of India and Abroad. In: Library and Information Technology: in pursuit of excellence. Papers presented in 38th All India Library conference, Bhubaneswar, 21-24 November, 188-195.

Geetha,V. (2012). Uses of e-resources by postgraduate students of the Department of Library and Information Science in Bishop Heber College, Trichy. In dynamic of librarianship in the knowledge society.Vol.2.p.551-558.

Gomez,R. & Kemly,C(2011),Users of ICT at public access centers:Age,education,gender and income differences in users. International Journal on Information Communication Technologies and Human Development, 3(1),1-20.

Krishna, R.(2011). Information-seeking behaviour of humanities researchers in the Internet age.In national seminar on collection development. Silchar: Assam university,Feb.10-11.

Mishra, R.N and Gohain.R.R. (2010). Use and usage statistics of electronic resources at central library, Tezpur University: a case study. In Re-Engineering Of Library And Information Services In Digital Era,7th convention PLANNER, Ahmedabad: INFLIBNET.p.183-205.

Natarajan,N.O.(2012).Use of e-resources by the faculty of education in Annamalai University-A study.In Dynamic Of Librarianship In The Knowledge Society.Vol.2.p.524-528.

Patra, M.(2012). Information access and management in digital era. New Delhi: SSDN Publishers.

Rajeswari,D. (2005). Use of information sources in digital environment : A case study. In Multilingual Computing And Information Management In Networked Digital Environment. proceedings of the 3rd International CALIBER, Ahmedabad: INFLIBNET. p.705-711.

Sarmah,M.(2011). IT applications in academic libraries of Assam. Guwahati: EBH Publication.

Sharma, N.N.(1994).College librarians of Assam: How they can face the challenge of the time. In: A. Buragohain and K.N. Baishya (Eds.), Librarianship And Information Technology. New Delhi:Omsons,130-134.

Umesh, K.Y. (2012.) Best practices adopted in academic libraries and information resource centres. International Journal of Information Dissemination and Technology, 2 (3), 186-188.

STEGANOGRAPHY
Art Of Hiding Information

Geeta Wahi
Mukand Lal National College, Yamuna Nagar

A B S T R A C T

A lot of applications are Internet-based and it is important that communication should be made secret and due to increase in the number of unauthorized access recorded during exchange of information between the source and destination has need for a more robust method for securing data transfer. Steganography andCryptography are important techniques that change information in to hide their presence. Steganography proves as abetter way of securing message than cryptography which onlyfocuses on the content of the message not the existence of the message. This paper will take an in-depth look at this technology byintroducing the reader to various concepts of Steganography, a briefhistory of Steganography.

Keywords—Steganography· History · Difference With Cryptography· Text steganography · Image steganography

INTRODUCTION

The word steganography combines the Ancient Greek words steganos, meaning "covered, concealed, or protected", and graphein meaning "writing". The first recorded use of the term was in 1499 by Johannes Trithemius in his Steganographia. In stegnography secret message does not attract attention to itself as an object of scrutiny.Steganographyis the art and science of writinghidden messages so that none but sender andrecipient realize there is a hidden message steganography is that it can be used to secretlytransmit messages without the fact of the transmission beingdiscovered. The main goal or steganography is to communicate securely in a completely undetectable manner and to avoid drawing suspicion to the transmission of a hidden data. It is ancient art of hidingmessages so that they are not detectable.

EXAMPLE

Susan eats truffles. Under pressure, that helps everything before 0wning Major Bullwinkle.
Set Up the bomb

HISTORY OF STEGANOGRAPHY

Through out history Steganography has been used to secretly communicate information betweenpeople.

Some examples of use of Steganography in past times are:

1. During World War 2 invisible ink was used to write information on pieces of paper so that thepaper appeared to the average person as just being blank pieces of paper.Liquids such as urine, milk, vinegar and fruit juices were used, because when each one of thesesubstances is heated they darken and become visible to the human eye.
2. In Ancient Greece they used to select messengers and shave their head, they would then writea message on their head. Once the message had been written the hair was allowed to grow back.After the hair grew back the messenger was sent to deliver the message, the recipient wouldshave off the messengers hair to see the secret message.
3. Another method used in Greece was where someone would peel wax off a tablet that wascovered in wax, write a message underneath the wax then re-apply the wax.The recipient of the message would simply remove the wax from the tablet to view the message.

DIFFERENCE IN STEGANOGRAPHY AND CRYPTOGRAPHY

Cryptography is the study of hiding information cryptography focuses on keeping the contents of a message secret, in cryptography message may be converted in the form of codes where each character is substitude by alternate one steganography focuses on keeping the existence of a message secret. Message can be hide by using different covers Steganography deals with composing hidden messages so that only the sender and the receiver know that the message even exists.Steganography does not alter the structure of message. Whereas cryptography alters the structure of message.In Cryptography data is made unreadable by the third party alters, without hiding. In Steganography it appears that no information is hidden at all .no one will attempt to decrypt the info hide, without altering.

STEGANOGRAPHY IN TEXT

Text steganography can be achieved by altering the text formatting, or by altering certain characteristics of textual elements (e.g., characters). Various Text Steganography methods are explain below:

LINE SHIFT CODING

In Line Shift Coding method, text lines are vertically shifted to encode the document uniquely. Encoding and decoding can generally be applied either to the format file of a document, or the bitmap of a page image. By moving every second line of document either 1/300 of an inch up or down, it was found

that line-shift coding worked particularly well, and documents could still be completely decoded, even after the tenthphotocopy.But in this method, the reader can see by usingdistance determining tools or if he copies part of the text, the information is missed. This method is good for printed texts because the text must be scanned for copying it, its line must be transferred again to the elsewhere and reprinted.

WORD SHIFT CODING

In word-shift coding, codewords are coded into a document by shifting the horizontallocations of words within text lines, while maintaining a natural spacing appearance.This method can also be implemented both in the text file and its picture.

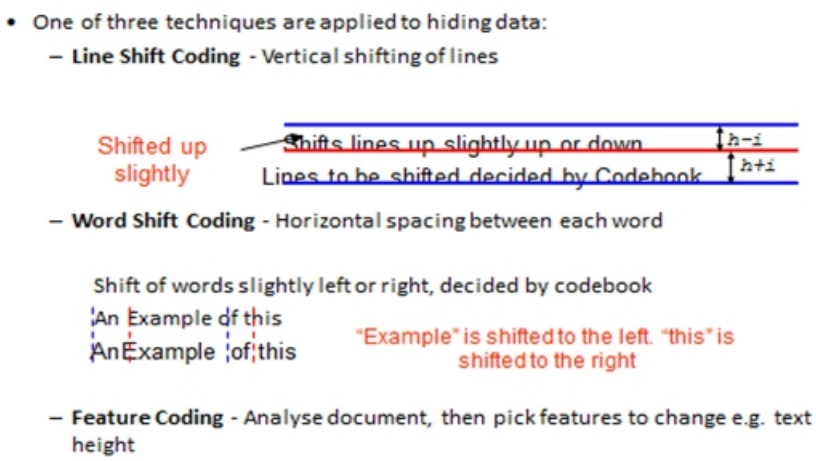
Example

How word-shifting might work. For each text-line,the largest and smallest spaces between words are found. The largest spacing isreduced by a certain amount, and the smallest is extended by the same amount. This maintainsthe line length, and produces little visible change to the text.

This method is seen less by the reader because changing the distance between words for filling one line is very common but word shift coding is observable and identifiable by two methods. Distance Algorithm:In this algorithm current text can be compared and find hidden information. The second method is point to point reviewing of picture text to specify the changed distances. This method is very time consuming and the possibility of finding hidden information is very high in it.

FEATURE CODING

This is a coding method that is applied either to documentation file or picture features are altered, or not altered, depending on the codeword. Decodingrequires the original image, or more specifically, a specification of the change in pixels at a feature. There are manypossible choices of text features; here, we choose to alter upward, vertical endlines - that is the tops of letters, a, b, h, etc.These endlines are altered by extending or shortening their lengths by one (or more) pixels, but otherwise not changingthe endline feature. In this method, because of the existence of many places for changing in the text, much information can be hidden, while the reader pays no attention to them at all.



SEMANTIC METHOD

A text can be placed and hidden in another text by using synonym word, for example the word Complicated can have the value of 1 and Difficult can equal to 0. Sometimes, this method makes the text meaningless.

SYNTATIC METHOD

In syntactic method, some dots or text signs is used. It is a very attractive method Syntactic method is almost performable in English language, this method has low efficiency, which means that for hiding only some bits, the main text equals many kilobits is needed.

STEGANOGRAPHY INIMAGE

It is use to hide a message inside an image without changing its visible properties.

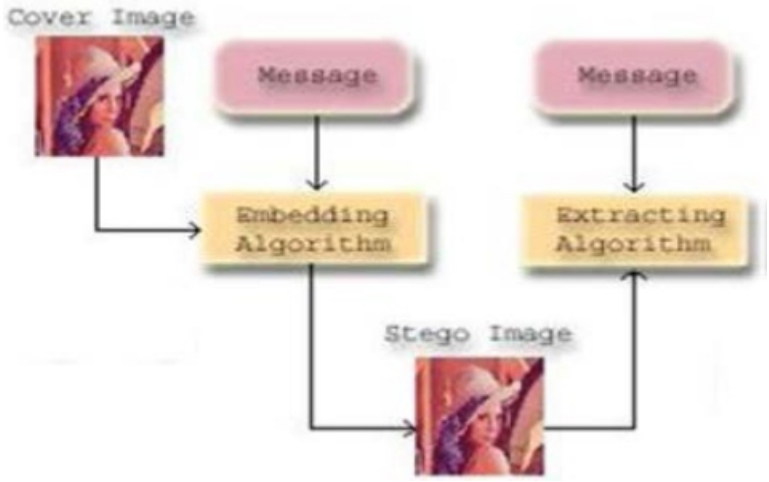
Image steganography terminologies are as follows:-

Cover-Image: Original image which is used as a carrier for hidden information.

Message: Actual information which is used to hide into images. Message could be a plain text or some other image.

Stego-Image: After embedding message into cover image is known as stego-image.

Stego-Key: A key is used for embedding or extracting the messages from cover-images and stego-images.



Least Significant Bits

A simple approach for embedding information in cover image is using Least Significant Bits (LSB). The simplest steganography techniques embed the bits of the message directly into least significant bit plane of the cover image in a deterministic sequence

When using a 24-bit color image, a bit of each of the red, green and blue color components can be used, so a total of 3 bits can be stored in each pixel

For example, the following grid can be considered as 3 pixels of a 24-bit color image, using 9 bytes of memory:

(001001111110100111001001)

(001001101100100011101000)

(110010010010011111101001

When the character A, which binary value equals 111000111, is inserted, the following grid results:

(00100111 1110100111001001)

(00100110 11001000 11101000)

(1100100100100111 11101001)

Masking and Filtering

It manipulates the luminance of particular areas of the image in order to encode dataThese techniques hide information by marking an image, in the same way as to paper watermarks. These techniques embed the information in the more significant areas than just hiding it into the noise level. The hidden message is more integral to the cover image. Watermarking techniques can be applied without the fear of image destructiondue to lossy compression as they are more integrated into the image.

REFRENCES

www.ijcaonline.org/volume9/number7/pxc3871887.pdf
http://www.academia.edu/8084609/IMAGE_STEGANOGRAPHY
www.liacs.nl/home/ tmoerl/privtech.pdf
http://www.seralliance.com/enews/vol2no1/pdfs/steganography_jan05.pdf
http://repository.root-me.org/St%C3%A9ganographie/EN%20-%20Image%20Steganography%20Overview.pdf

Development Of Predictive Simulation Models For Drug Dissolution Parameters Computing

Dr. H. B. Bhadka
Dean, Faculty of Computer Science,
C. U. Shah Univeristy, Wadhwan. Gujarat (India)

ABSTRACT

Over recent years, drug dissolution computing has been the subject of intense and profitable scientific developments. Whenever a new batch profile is developed or produced, it is necessary to ensure that drug dissolution occurs in an appropriate manner. The quantitative analysis of the values obtained in dissolution tests is easier when statistical formulas that express the dissolution results as a function of some of the dosage forms parameters are used. In most of the cases the theoretical concept does not exist and some empirical equations have proved to be more appropriate.

Keywords—Drug Dissolution, Drug release; Drug release models simulation; Parameters Computing.

1. INTRODUCTION

Dissolution testing plays an important role in pharmaceutical quality control and in the development of solid, semi-solid, and transdermal pharmaceutical forms. The dissolution kinetic is reexamined under simulated physiological conditions, which are specified in both the U.S. Pharmacopeia (USP) and the European Pharmacopeia (EP) dissolution testing regulations. As such, these analytics are performed in a highly regulated Good Manufacturing Practice (GMP) environment, and present particular challenges for that facilitate those computations software applications. The role of computer based information systems has considerably increased in pharmacy as well as clinical practice in the last decade. However, the use of such systems in dissolution test is still not widespread. Several mathematical systems are commercially available for dissolution parameters calculations. In addition, comprehensive systems customized to specific needs have also been developed. The use of such systems in dissolution parameters calculations, questions regarding the role of structured data versus free text input, standardization of nomenclature, and compatibility with other systems, are hotly debated. This paper has in its scope the above stated considerations in the development of software for dissolution parameters calculations records, which attempts to resolve some of these issues. The model described herein is specifically designed to meet the requirements of the dissolution parameters calculations of a

tertiary referral. An additional module has been included to allow modification and update of previously recorded data. A unique number assigned to each has been used as a primary identifier throughout the record.

2. MATERIALS AND METHODS

Requirements:

The central objective of this initiative is to create a data model capable of accurately representing the calculations for dissolution parameters in a computer-suitable format. The main requirements include that the system should (a) be simple enough to be directly operated by the analytical scientist(s) in analytical research laboratory, (b) can easily run on personal computers,(c) allow comprehensive data entry conforming to accepted procedures which are currently carried out, (d) can generate a printed report, (e) allows modification and update of data, and (f) permit subsequent statistical analysis of records in a tabular format and displays the results of analysis. As dissolution data are to be handled by analytical scientists with minimal previous computer experience, an emphasis is laid on a user-friendly interface.

Software Construction:

The software has been developed in visual logway in Visual Basic. It has a set of two screens for data entry. One relating to (a) calibration curve and the other relates to (b) cumulative percentage release. Data flow is designed in two directions: (a) to a database, after appropriate coding, for storage and subsequent analysis at a later date, and (b) to the report generator. An
additional module is included to allow modification and update of previously recorded data. Another module is designed for filling test reports on specimens obtained during the procedure, as and when these results became available. A unique identifier assigned to each test record is to be used as a primary identifier throughout the record.

Data Entry:

Modules are developed to allow easy user access and facilitate data entry. On completion of one module, automatic transfer to the subsequent module is envisaged. The basic module is structured as a large window, with smaller sub-windows appearing only on demand. The entire software is menu driven, with a simple and consistent hierarchical structure. As far as possible, all fields are structured, with

the user allowed to choose one or more options from list of choices. These options included important and/or commonly observed conditions, and are chosen to cover majority of everyday findings after consulting experienced faculty members and reviewing previous records. A standard terminology developed for the structured items based on available literature and general consensus. The fixed choices are displayed either as searchable list boxes, check boxes, or as radio buttons. Free text is allowed in some fields, such as the information beyond the fixed choices available to the user. To allow complete data acquisition in each test, all data fields are marked mandatory, and the user is not allowed to proceed to a subsequent field without recording data in such fields. (Fig. 1 and 2)

Debugging and Modification:

After initial development, the software is tested over a four-week period by input of data. An attempt is made to rectify problems faced initially by the users. Opinion is sought from faculty members regarding possible modifications and improvements. Inconsistencies in the programming script, which gave rise to error messages during operation of software, are corrected. Finally the software is put to routine use.

Software Validation:

To evaluate the actual utility of the software, all consecutive test records entered using this computer software. Analytical scientists are asked to assess the overall quality of the reports and the content of information. After entry of data for 60 consecutive test procedures, these details are subjected to statistical analysis to evaluate the robustness of the database component.

Linearity or calibration curve [6]:

The linearity of an analytical procedure is its ability (within a given range) to obtain test results, which are directly proportional to the concentration (amount) of analyze in the sample. A linear relationship should be evaluated across the range of the analytical procedure. It may be demonstrated directly on the drug substance (by dilution of a standard stock solution) and/or separate weighing of synthetic mixtures of the drug product components, using the proposed procedure. The latter aspect can be studied during investigation of the range.

Linearity should be evaluated by visual inspection of a plot of signals as a function of analyze concentration or content. If there is a linear relationship, test results should be evaluated by appropriate statistical methods, for example, by calculation of a regression line by the method of least squares.

In some cases, to obtain linearity between assays and sample concentrations, the test data may need to be subjected to a mathematical transformation prior to the regression analysis. Data from the regression line itself may be helpful to provide mathematical estimates of the degree of linearity. The correlation coefficient, y-intercept, slope of the regression line and residual sum of squares should be submitted for regulatory purpose. A plot of the data should be included. In addition, an analysis of the deviation of the actual data points from the regression line may also be helpful for evaluating linearity. For the establishment of linearity, a minimum of 5 concentrations is recommended. For the dissolution, concentrations of drug are calculated from the respective calibration curve (Fig. 3).

Dissolution study [1, 4-5]:

In vitro dissolution specifications are established to guarantee batch-to-batch consistency and to indicate potential bioavailability problems. For new drug products, dissolution specifications must be based on data obtained from the batch used in the bioavailability assay (bio-batch). For generic drugs, the dissolution specifications are generally the same of the reference drug product. These specifications are confirmed by testing the performance of the bio-batch dissolution. If the generic drug dissolution is substantially different from the reference drug product dissolution, and the in vivo study had proved the bio-equivalence between them, a different dissolution specification for the generic drug can be established, provided it is based upon a validated IVIVC. In that case, the specification must be fulfilled throughout the permanence of the generic drug in the market. The specifications must be based on the bio-batch dissolution characteristics. If the formulation developed for commercialization differs significantly from the bio-batch, the comparison of the dissolution profiles and the bio-equivalence study between these two formulations is recommended.

The dissolution tests must be undertaken under such conditions as: basket method at 50/100 rpm or paddle method at 50/75/100 rpm. To generate a dissolution profile, at least five sampling points must be obtained of which a minimum of three must correspond to percentage values of dissolved drug lower than 65% (when possible) and the last point must be relative to a sample period of time equal to, at least, the double of the former period of time. For drug products of rapid dissolution, samples at shorter intervals (5 or 10 minutes) may be necessary. For drug products with highly drug lower than 65% (when possible) and the last point must be relative to a sample period of time equal to, at least, the double of the former period of time. For drug products of rapid dissolution, samples at shorter intervals (5 or 10 minutes) may be necessary. For drug products with highly soluble drugs that present rapid dissolution (cases I and III of BCS), a dissolution test of a single point (60 minutes or less) that proves a dissolution of, at least, 85% is sufficient for batch to batch uniformity control. For drug products containing drugs poorly soluble in water, which dissolve very slowly (case II of BCS), a two points dissolution test, that

that is, one at 15 minutes and another at 30, 45 or 60 minutes, to ensure 85% of dissolution is recommended (Fig. 4 and 5).

Dissolution Efficiency [7]:

Khan suggested Dissolution Efficiency (D.E.) as a suitable parameter for the evaluation of in vitro dissolution data. D.E. is defined as the area under dissolution curve up to a certain time „t" expressed as percentage of the area of the rectangle described by 100% dissolution in the same time. The D.E. values are calculated from the dissolution data. (Fig. 6)

$$\text{Dissolution efficiency (D.E.)} - \frac{0\int^{t} y.dt}{y100^{t}} x100$$

Comparison of dissolution profiles by similarity and dissimilarity factor [2-3, 8-9]:

To avoid the requirement of bioequivalence studies of the immediate release pharmaceutical forms of lower dosage, when several presentations with the same formulation exist, the dissolution profiles must be compared and must be identical among all dosages.

Until recently, single point dissolution tests and specifications have been employed to evaluate scale-up and post-registration changes. When minor alterations are carried out, the single point dissolution test may be adequate to ensure drug product quality and performance. For majoralterations, the comparison of dissolution profiles obtained in identical conditions between the altered formulation and original one, is recommended. In this comparison, the curve is considered as a whole, in addition to each sampling point of the dissolution media, by means of independent model and dependent model methods. Independent model method employing the similarity factor. A simple independent model method employs a difference factor (f1, Fig. 7) and a similarity factor (f2, Fig. 8) to compare dissolution profiles. Factor f1 calculates the percentage difference between two the profiles at each sampling point and corresponds to a relative error measure between the profiles:

$$f_1 = \{\left[\sum^{n}{}_{t-1} | R_t |\right] + \left[\sum^{n}{}_{t-1} R_t\right]\} x100$$

where:

n = number of sampling points

Rt = value dissolved in time t (percentage), obtained with the reference product or with the original formulation (before the alteration)

Tt = percentage value dissolved from the altered formulation, in time t.

Factor f2 corresponds to a similarity measure between the two curves:

$$f_2 = 50 \times \log\left\{\left[1 + \frac{1}{n}\sum_{t=1}^{n}\left|R_t - T_t\right|^2\right]^{-0.5} \times 100\right\}$$

The procedure is described as follows:

- Determine the dissolution profile of products, test and reference, using twelve units of each.
- Calculate factors f1 and f2 using the equations presented previously.
- Criteria for two dissolution profiles to be considered similar.
- The nominal range of f1 and f2 values are 0 to 15 and 50 to 100, respectively.

3. RESULTS

A software for drug dissolution parameter computation developed with Graphical User Interface (GUI). During execution, it takes for the drug concentration, instrument response and time data.

After taking input it display list where user can opt for specific set of computations and can get the results for desired set of computation. The software supplements visualization along with computation. The user can opt for reports to be provided by the software. It generate calibration curve, cumulative percentage release, dissolution efficiency, comparisons of two products through similarity and dissimilarity factors. The software has various modules for input and modification of data, computation of various parameters and visualization with facilities to generate reports of dissolution parameters. The use of interface is designed for work with much ease in respecting. With little practice, scientists soon became adept at entering details correctly and quickly. The slightly increased time of data entry into the computer is more than made up by uniform and complete report generation. A user-friendly software providing computation and visualization parse drug dissolution parameters. The analytical scientists can utilize the software for intensive research as wide variety of parameter computation at simple key stroke.

The computer software currently used has two modules for data input: (a) calibration curve, and (b) cumulative percentage release. The data is linked to a MS-SQL Server having a set of two tables related to (a) calibration curve, and (b) cumulative percentage release and their reports. The two tables are linked to each other using the unique number. Another module deals with screen preview of reports and generation of printed reports.

In the calibration curve module, the number identifies each test record uniquely. The date of procedure is

automatically derived from the system date maintained by the computer clock, but can be changed manually. The user has to enter the number of observations of concentration and instrumental response. After completion of calibration curve test record, the user is transferred directly to the „cumulative percentage release" module. The possible locations in the dissolution parameters tree are represented by a cascading hierarchy of tables. An additional table listing the appropriate divisions/segments appears.

On completion of data entry, the user is transferred to the print module, where he can preview the report prior to printing. The printed report contains all the information entered in the database. It also contains a standard set post- procedure instruction for the test, and also has space for signatures for the analytical scientist carrying out the procedure.

Problems initially faced by users are primarily related to data entry. Scientists, not having any working knowledge of computers, encountered problems such as a slow speed of data entry and failure to enter data in mandatory fields (with a consequent error message that did not allow the user to proceed further without rectifying the mistake). With little practice, they became adept at entering details correctly and quickly. Almost all the analytical scientists reported a slightly increased time of data entry into the computer, in comparison to writing reports on a standard printed proforma. However, all agreed that the report and data generated through the software are uniformly complete, and more than made up for the extra time spent. The new report has a uniform and easily understood structure, and is free of any inadvertent omissions.

The database component is evaluated by analyzing 60 consecutive records entered over a 4-month period. Data access and analysis are easily and quickly performed. Data are found to have been completely transferred from data entry screens to the database and no missing values are encountered.

4. DISCUSSION

Structured input and free-text input represent two fundamentally different ways of entering data into a computer. Initial reports of test databases relied heavily on text based tools. Such input facilitates personalized style and flexibility in description of test records, and generates a well readable report. However, free-text input weakens the utility of the database, as it is not suited to subsequent analysis. Structured input and the resulting categorical data offer an important advantage in this regard. Data thus entered is more likely to be complete and is well suited for research and analysis, as well as for the generation of analytical reports and for quality control. It has been estimated that use of computerized test records improves completeness of data entry by more than 50 percent. However, a major trade-off

for structure is flexibility. We therefore used a basic structured data entry protocol, supplemented by use of free text only under special situations. Besides operator related factors, it is related to the amount of free text entered and the number of tables accessed during structured data entry. However, the additional effort is rewarded by a more comprehensive, precise and accurately documented report.

A major feature of the software is the powerful database component. This portion of the software has been built as a set of two interrelated database in MS-SQL Server, which can easily handle large database and also offers a wide range of analytical tools through a versatile query system. We have evaluated the robustness of this module of the software through an analysis of 60 consecutive test records.

Although such analysis requires some working knowledge of the database system, it is easy of learn. No data is lost and statistical analysis could be easily performed. Both user-friendliness of the software and completeness of data entry are critical to the success and acceptance of such software. This software allows easy integration of buttons, text boxes, check boxes and fields for free text to achieve this end. The format for data input is optimized through continuous interaction between scientists and the programmer. Scientists and other faculties are involved early and frequently during the development of the software, so that they are able to contribute ideas and advice. The software has been under routine use, and has performed well in areas of data entry, report generation and data analysis. Successful development and routine application of the database is, however, only a short-term achievement. The system is adaptable and capable of keeping pace with new technological advances.

REFERENCE:

1. *"Dissolution testing", United State Pharmacopoeia, XXIII, NF XVIII, The USP convention, Inc, Rockville, MD, (1995).*
2. *M. C. Gohel and M. K. Panchal, "Refinement of lower acceptance value of the similarity factor f2 in comparison of dissolution profile". Dissolution Technol., vol. 2, pp. 18-22, 2003.*
3. *M. C. Gohel, K. G. Sarvaiya, N. R. Mehta, C. D. Soni, V. U. Vyas and R. K. Dave, "Assessment of similarity factor using different weighting approaches". Dissolution Technol., vol. 11, pp. 22-27, 2005.*
4. *Guidance for industry: Dissolution testing of immediate release solid oral dosage forms. US food and drug administration, Rockville, MD, USA, 1997.*
5. *Guidance for industry: Immediate release solid oral dosage forms, scale-up and post- approval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation. US food and drug administration, Rockville, MD, USA, 1995.*
6. *International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use, ICH harmonised tripartite guideline validation of analytical procedures: Text and methodology Q2 (R1), 2005.*
7. *K. A. Khan and C. T. Rhodes, "Concept of dissolution efficiency". J. Pharm. Pharmacol, vol. 27, pp. 48-49, 1975.*
8. *J. W. Moore and H. H. Flanner, "Mathematical comparison of dissolution profiles". Pharm. Technol., vol. 20, pp. 64-74, 1996.*
9. *J. E. Polli, G. S. Rekhi, L. L. Augsburger and V. P. Shah, "Methods to compare dissolution profiles and a rationale for wide dissolution specification for metoprolol tartrate tablets", J. Pharm. Sci., vol. 86(6), pp. 690-700, 1997.*

Fig. 1 Input form of calibration curve – Data entry

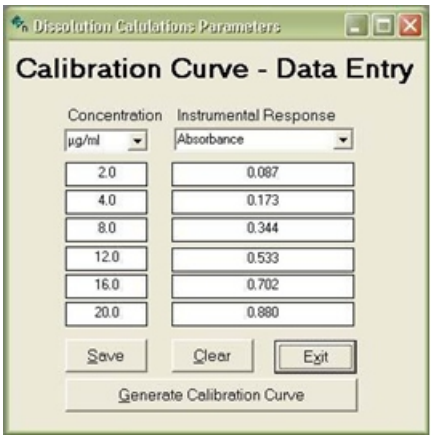


Fig. 2 Input form of Cumulative percentage release – Data entry

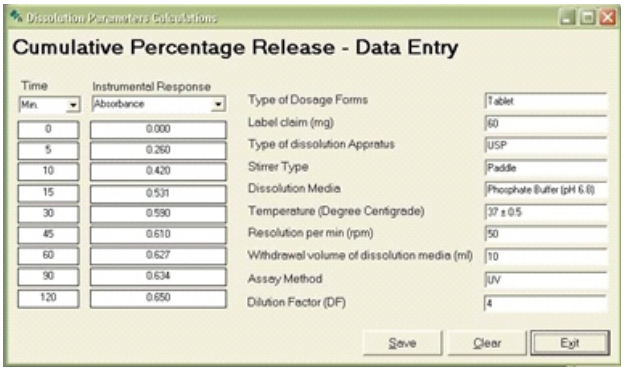


Fig. 5 Output report of cumulative percentage release

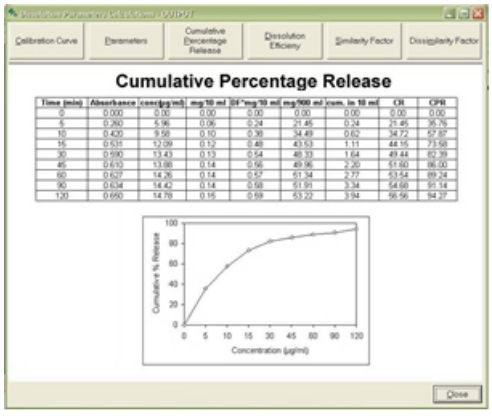


Fig. 3 Output report of calibration curve

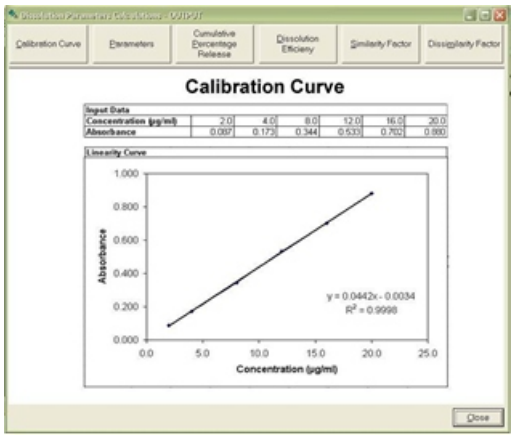


Fig. 4 Output report of parameters

Fig. 7 Output report of similarity factor

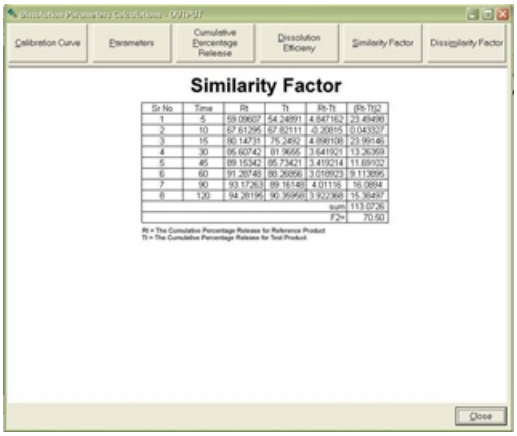


Fig. 6 Output report of dissolution efficiency

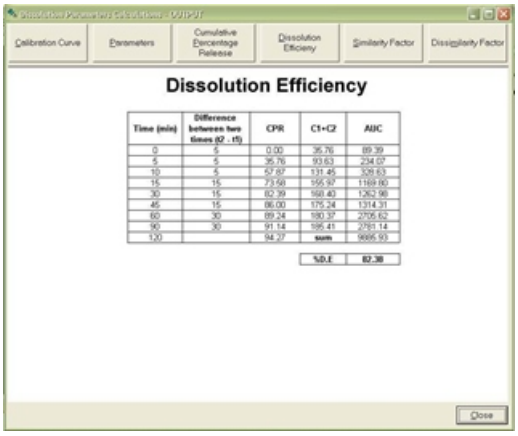
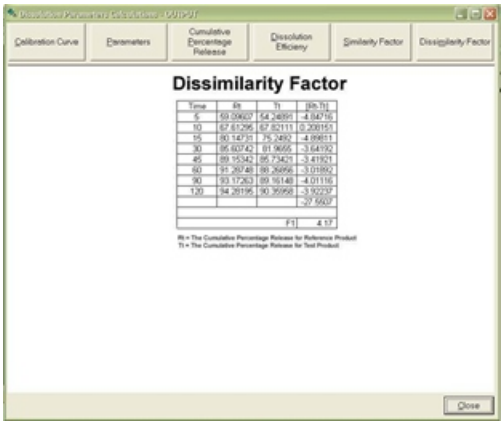


Fig. 8 Output report of dissimilarity factor



Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

-

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005