UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294)

A Metadata Approach to Context Development for Big Data

Nitin Varma[1] and Pradip Kumar Bala[2]
[1]Analytics & Information Systems, Indian Institute of Management Ranchi, India
[2]Analytics & Information Systems, Indian Institute of Management Ranchi, India

Abstract - Since most of business value is believed to be locked away unstructured, and most of it so far textual, this elevates "text mining" to a pedestal in the Big Data world. It, however, also brings forward the bitter realization- that unlike mining structured data, there is little use of "direct" mining that may be possible on unstructured data. In fact, the current approaches in Big Data analytics are already under attack, one of the often-quoted reasons being lack of context, even though there is agreement that the concept of Big Data continues to hold promise. Recent experiences and surveys with industry leaders and practitioners have shown that the Big Data analytics problems may be better solved through development of context, which in turn may be possible also via metadata. Despite the importance of meta data to real life Big Data analytics, extant academic literature lacks in clarifying the role and importance of meta data for context development, no existing work is able to provide as comprehensive an understanding. Therefore, this research builds a case for meta data for context development- so that big data analytics may be able to deliver increasingly towards its true potential. It utilizes feedback and knowledge from industry to build this foundation in academics. It, however, does not stop at that, since corresponding examples are practically hard to find. The position of this paper is demonstrated through three text data case examples - how metadata can help in developing context to explain aspects of business problems, so that interested researchers can develop a comprehensive understanding.

Keywords: ***Big Data, Meta Data, context, metadata, context-aware.***

I. INTRODUCTION

The term "Text Analytics" describes a set of linguistic, statistical and machine learning techniques that model and structure the information content of textual sources for various purposes (Hobbs, 1982) – research, analysis and/or intelligence – in public and private domains.

It may be generally accepted that upto 80% of business knowledge is locked in unstructured text (Adrian, 2011). Therefore, potentially tremendous business knowledge could be uncovered using "Text Analytics".

Declaring that Text Analytics, a sub-set of Big Data Analytics, has so much potential as to be the next "big" thing, the Time Magazine chose to devote its cover page and cover story (Belsky, 2012) to the theme: "Why Text Mining May Be The Next Big Thing".

Realizing the importance of Text Analytics amidst this Big Data revolution, the U.K. (Britain) became the first country in the world, to fund, start-up and run a Government sponsored institute dedicated only to text mining, called NaCTeM (NACTEM, 2013).

Manyika (2011) states McKinsey defined Big Data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

Other definitions for Big Data (Grimes, 2013) have attempted to expand beyond the original three V's – volume, velocity and variety to seven V's: volume, velocity, variety, veracity, variability, value and visualization. "Viability" is being said to be the eighth "V" - since some Big Data scientists have now come to an understanding that at times perfecting only 5% of the Big Data variables that are also viable, may deliver justifiable value.

Lack of clarity in definition apart, challenges in the Big Data world are now beginning to emerge. According to the Analytics Magazine, despite Advanced Analytics being the fastest growing segment within IT industry [as per Gartner survey, 2012-13], how to best use analytics on Big Data still begets confusion (Analytics Magazine, 2014):

"... the recently released (Gartner) report showed that advanced analytics grew the fastest (12.5 percent increase in revenue from 2012-2013) of all the sub-segments surveyed …, …, confusion still reigns around how to best leverage analytics on Big Data".

II. RESEARCH PROBLEM AND METHODOLOGY

As discussed above, issues were reported in implementation of Big Data - that relate to the importance of Context and Metadata. However, as interest developed in this topic, it was found that there is a huge research gap in academic literature. For example, this work considered literature available at Emeraldinsight.com and found not even a single paper that links context to meta data for Big Data and no work provides even a foundational article for research. It is the absence of literature that prompted this research.

Therefore, the methodology for this work required considering sources that could provide valuable information. This work draws upon feedback from industry, also upon surveys conducted by a diverse set of industry leaders, from experiences of industry experts and from practices adopted by those actually carrying out metadata and Big Data analytics in various application areas. This work does not shy from utilizing academic contributions, when these were available.

III. FINDINGS

This section presents a view of context and meta-data issues as reported in the domain of Big Data. Also, from the emphasis developed, three cases are presented to develop a strong foundation for comprehensive understanding.

3.1. Experts Contribution: Building the Emphasis

SAS, the world's leading data-mining software maker has shared five challenges in Big Data Analytics, identifying once again the need to understand data in the right context (SAS, 2013): (1) meeting the need for speed (2) understanding data in the right context (3) ensuring data quality (4) being able to visualize output in a meaningful manner and (5) dealing with outliers.

The New York Times, quoting experts and spilling the beans of Big Data discord in public domain, has reported nine issues related to Big Data including issues about combining (Gary and Davisparil, 2014) data from various sources and thus resultant ill effects, ostensibly due to mixing of contexts, if any.

In a scathing attack on Big Data Analytics, the Financial Times London has quoted statisticians and industry cases – to point out how lack of understanding (context), for example of Twitter data, can lead to terribly skewed results (Harford, 2014).

Big Data research, like all empirical research needs to stand on a foundation of measurement. A major question that extends therefore to Big Data also, is: is the instrumentation actually capturing the theoretical construct of interest (Lazer, Ryan, King and Vespignani, 2014)- is context being captured in current Big Data scenarios?

The Economist explains some of the above and many more disgruntled voices as necessary hype-busting, while still maintaining that the concept of Big Data analytics holds value (K.N.C., 2014).

Therefore, the current state of Big Data and the assumptions lurking in it are well summarized by the OCCAM framework floated by Kaiser Fung on the Harvard Blog Network (Fung, 2014) that again points towards lack of capture, loss or mixing of contexts in current practice. As per the OCCAM framework, Big data is: (1) Observational: much of the new data come from sensors or tracking devices that monitor continuously and indiscriminately without pre-specified objectives (2) Lacking Controls: controls are typically unavailable, making valid comparisons and analysis more difficult (3) Seemingly Complete:the availability of data for most measurable units and the sheer volume of data generated is unprecedented but still only seemingly complete (4) Adapted:third party data is likely not aligned to data scientists objectives (5) Merged: different datasets are combined, exacerbating the problems relating to lack of definition and misaligned objectives.

The above issues are on top of the already recognized complexity of encoding and decoding from text. It has, for a long time been known that text structures depend crucially on what we regard as "common-sense" knowledge, which despite-or, more likely, because of- its everyday nature is exceptionally hard to encode and utilize in algorithmic form (Witten, 2005).

Also, Big Data – its being unstructured, hardly lends to its utility for querying, in its native form. This is well posited by Bill Inmon (Inmon, 2013):

"...there are indeed some important simple differences between data. And why are those simple differences important? They are important because everyone is talking about big data today. You know, the kind of data found in Hadoop. Does anyone stop and realize that all data in Hadoop is unstructured and that you can do only the most basic of queries against that data? Something to think about."

Making matters even worse, a survey of data scientists published by CIO Magazine Survey (Olavsrud, 2014), brought out the frustrations of data scientists. At least two of the problem areas identified are - linking various data sources and putting together data from various sources.

While now it is established that current Big Data is context-challenged, the two key expectations - to the contrary, from Big Data, as per three consecutive annual IBMsurveys of Chief Executive Officers and Chief Marketing Officers, are that (IBM, 2014):

- Big Data will deliver the empowered customer by enabling named customer-centric objectives such as improving customer satisfaction and reducing churn

- Big data will help understand and predict customer behavior by "creating" a complete picture (i.e. context) of customers' preferences and demands Master data can come to rescue and help in building context for delivering customer-centric outcomes. With Master Data Management (MDM), organizations can correlate unstructured text to existing master records, discover linkages between text and relevant master entities, and enrich the master record with additional information. As the IBM graphic (IBM, 2014) below makes it clear - master data management creates context for Big Data, while Big Data creates context for master data management.

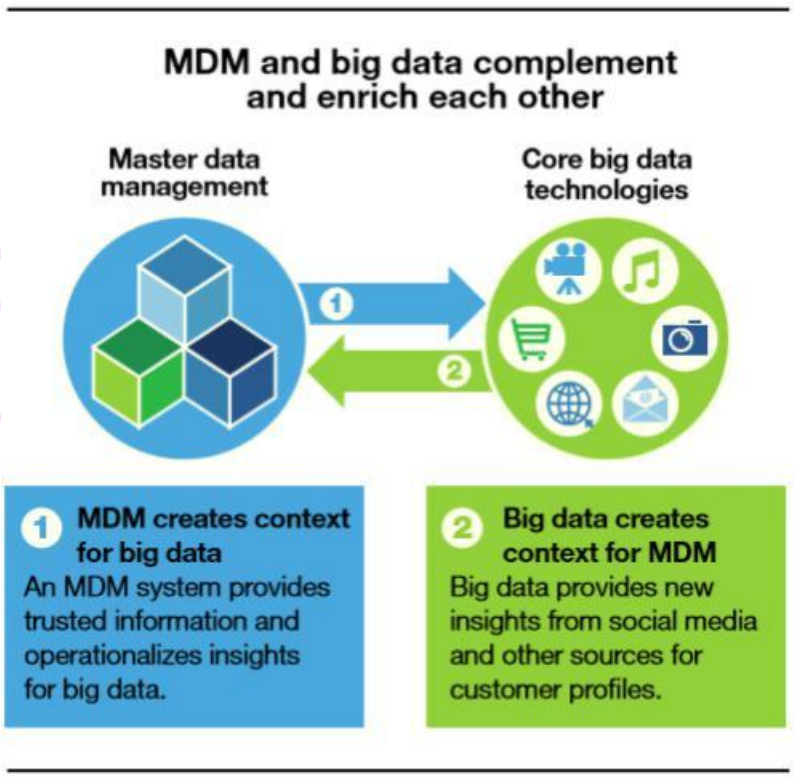


Fig1 The MDM Advantage: Creating Insight From Big Data[17]

Metadata and metadata management become even more important when dealing with large, complex, and often multi-sourced data sets (Prevosto and Marotta, 2013).

The documentation and description of datasets with metadata—data about data—enhances the discoverability and usability of data both for current and future applications, as well as forming a platform for the vital function of tracking data provenance (Dumbill, 2013).

As per McKinsey Global Institute (MGI) research, creating substantial new value does not necessarily require jumping directly to complex analytical big data levers (Manyika, 2014). MGI recommends techniques to "scrub" the data to remove errors and ensure data quality, to place data into standard forms, so that metadata can be added to describe the data being collected.

3.2 Developing Context: Three Case Examples

As per the Oxford (2014) dictionary, context may be defined as: "the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood". In this section three cases are utilized to provide a strong foundation for a comprehensive understanding.

Case Example 1: A recent M&A of Peter's Bank – which had presence in 50 countries in 100 cities, with ABCD Inc. - which had presence in 45 countries in 120 cities, has resulted in a vast number of offices for the merged entity: Peter's ABCD Bank in 72 countries worldwide. Metadata is available for each of the cities where pre-merger Peter's Bank and ABCD, Inc. had offices (sample GIS metadata for one city is shown below (MIT Geoweb GIS, 2014). Clearly metadata can help the new entity Peter's ABCD Bank develop the context to undertake necessary decision-making for maximizing convenience, services and benefits to the customers of the new entity, while minimizing commute times and overall operational costs (if that data is also made available). GIS metadata can help Peter's ABCD Bank optimize locations for maximum footfall, by consolidating and eliminating duplicate offices and by advising re-allocations that would enable low commute timings for employees.

Metadata_Date: 20170121

Metadata_ID_Scanner:
Employee_Card_Scanned: 45
Debit_Card_Scanned: 2430 (mandatory, to open main entrance door)

Metadata_Contact:
Contact_Information:
Contact_Organization_Primary:
Contact_Organization: ABCD, Inc. (ABCD)
Contact_Person: Data Team

Contact_Address:
Address_Type: mailing and physical address
Address: 380 New York Street
City: Redlands
State_or_Province: California
Postal_Code: 92373-8100
Country: USA

Contact_Voice_Telephone: XXX-XXX-XXXX
Contact_Facsimile_Telephone: XXX-XXX-XXXX
Contact_Electronic_Mail_Address: info@abcd.com
Hours_of_Service: 8:00 a.m.-5:30 p.m.

Time_Zone: Pacific time
Working_Days: Monday-Friday
Metadata_Standard_Version: NV-STD-001-2014
Metadata_Time_Convention: local time

Spatial_Domain (geo co-ordinates):
Bounding_Coordinates:
West_Bounding_Coordinate: -18.0000
East_Bounding_Coordinate: -25.0000
North_Bounding_Coordinate: 30.0000
South_Bounding_Coordinate: -55.9795

Thus, the above metadata (it may be analyzed in different ways using tools where necessary, for example to calculate distance from geo co-ordinates, etc.) may be sufficient to develop the context for the required decisions, and combined with more data (e.g. rent for each office), can aid development of context for even broader problem solving.

Case Example 2: Case of the Book in Abandoned Carts

On the web-site of the world's biggest online book-store visited by millions everyday, a book written by a certain author is suddenly appearing more often in abandoned carts than before, just as a new randomization algorithm for recommending books was implemented.

A dump of metadata related to the book in question is available as below and its resemblance to Dublin Core (NISO, 2014) is not just a co-incidence:

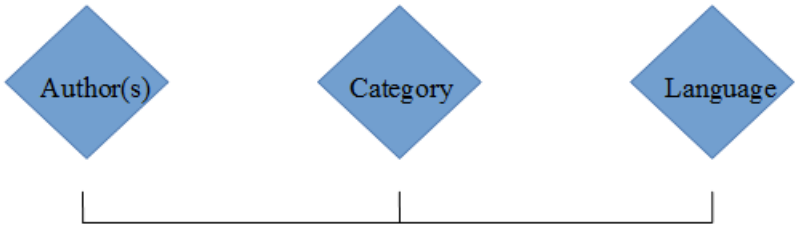
Title=”Metadata Demystified” Creator=”Brand, Amy” Creator=”Daly, Frank” Creator=”Meyers, Barbara” Subject=”metadata” Description=”Presents an overview of metadata conventions in publishing.” Publisher=”NISO Press” Publisher=”The Sheridan Press” Date=”2003-07-01" Type=”Text” Format=”application/pdf” Identifier=”http://www.niso.org/ standards/resources/ Metadata_Demystified.pdf” Language=”en”

Page_views_category: 13,456,987
Page_views_item: 6,998,798
Favorite_bookmarks_category: 24,983,832
Favorite_bookmarks_item: 16,235,982

It may be assumed that similar metadata for other objects displayed on the page is available.

Fig. 2 Metadata helps develop context to explain business situations:

CASE 2: why is the book getting abandoned in the shopping cart of late?



1. Is metadata information accurate? Or is some information missing and is critical?
2. Is there some related but extraneous information that is impacting buyers?
3. On what pages was the book displayed, what is the meta-data for those pages?
4. What other items were displayed on the page along with the book?
5. Which items appeared most frequently or least frequently along with the book?
6. Which items sold along with the book?
7. Is there a match between the number of searches, no. of in-category display and overall no. of displays of the book?
8. Is the category or the language itself under some kind of market-metamorphosis?

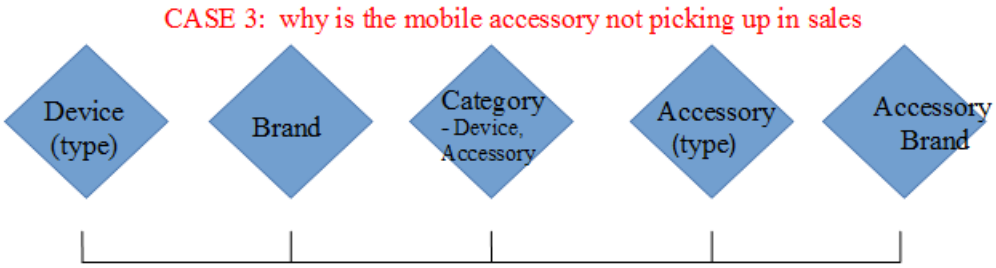
When the book was selling well	After the book abandonments became significant

The above metadata again may be sufficient to create a context to explain aspects of the situation: why is the book being abandoned?

Case Example 3: why is the mobile accessory not picking up in sales

When item x1 is purchased, some or all of the accessories xa1, xa2, xa3, xa4 etc. are also purchased. However, of late item xa3, which is a univeral phone cover, that goes with many 5.7" note-enabled phablets, is not picking-up sales despite six months of promotion.

Fig. 3 Metadata helps develop context to explain business situations:



1. Is metadata information missing, complete or accurate and is it critical?
2. Is some related extraneous information having an impact on prospects?
3. On what pages was the accessory displayed, what is the meta-data for those pages?
4. What other items were displayed on the page along with the accessory?
5. Which items appeared most frequently or least frequently along with the accessory?
6. Which items were sold along with the accessory?
7. Is there a match between the number of searches, in-category display and overall no. of displays of the accessory?
8. Is the market for the brand or category itself changing?

When the accessory was introduced	Despite 6-month promos for the accessory

Metadata similar to the book example in Case 2, as applied to the accessory in question above, may be sufficient to develop a context to explain aspects of the situation: why the accessory is not selling.

IV. IMPLICATIONS

This research discusses the role of meta-data and context in Big Data analytics and it emphasizes that these are far less understood in extant literature. This is evident, since most of the knowledge is drawn from industry experts and some, from the academic world.

This work brings forward the bitter truth- that unlike mining structured data, there is little use of "direct" mining that may be possible on unstructured data. It discusses findings from the Big Data experiences of even giants like Google, that failed to comprehend the Big Data issue in the right perspective – at least in the beginning, as is demonstrated through failure of Google Flu. In fact, the current approaches in Big Data analytics are already under serious attack, one of the often quoted reasons being lack of context, even though there is agreement that the concept of Big Data continues to hold promise.

Further, in this work, from recent experiences and surveys with industry leaders and practitioners, it is shown that the Big Data analytics problems may be better solved through building up of context, which in turn can be done via metadata. Findings from leading industry experts and organizations such as McKinsey are shared with the readers – to provide them a clear upfront view of the issues that can crop up if meta-data and context are not accorded their due place in the Big Data spectrum.

Also, this research demonstrates how meta-data can be used to achieve answers to business questions, through three cases. As a result of this research, future researchers – especially in academics, may be able to utilize this foundational work for being able to realize the full potential of Big Data.

It is only now that frustrations in making sense of Big Data have led data scientists to take another look at the role of context and therefore at the development of context through metadata. The hype of Big Data is now melting; bringing increasing numbers of scientists to the agreement that possible direct querying of Big Data is not going to unlock the true potential therein. This research brings to academia this "new" perspective on the role of context and the role of metadata in building context for Big Data analytics. Clearly, this perspective has been gathering momentum in the industry but is yet to find a major foothold in Big Data academia. It is expected that this work will join some of the early efforts worldwide in bringing into core academia, a distinct view on the importance of context and the importance of metadata for enhanced Big Data analytics.

In face of criticism of the current Big Data constructs by much respected popular press and even by venerated institutions like Harvard[11], even though there is agreement that there is value in the concept itself of Big Data, it is clear Big Data must deliver. For that to happen, brute force computation; machine or algorithm power has not been sufficient. This research clearly establishes that context development, through metadata holds the key to unlocking the promised value from Big Data analytics, and therefore also from Text Data analytics.

V. Limitations

The absolute shortage of academic literature in this area is an absolutely limiting factor. Though, there is much literature of various sorts available from industry and experts.

While through the case examples it is unequivocally demonstrated how context may be developed using metadata to approach and explain aspects of specific business problems, this paper makes no effort to elaborate how the required Big data and the required metadata may be extracted. There are technology intensive techniques, including extraction using XML or from separately maintained Big Data or metadata warehouses, that can be explored to serve that purpose.

This paper duly takes into cognizance the fact that context is required to be developed to explain specific a business situation, issue or problem. Metadata can be sufficient for developing context(s) to explain aspects of many business problems- however, combined with more data, should be able to explain even further. Thus it is upto practitioners to utilize additional data to further develop relevant context. For that reason, it becomes very challenging to work with context – because it can seem commonsensical to humans, but computers need to be told explicitly about it – in computing lingo.

Lastly, this paper stays abstains from discussion or classification of metadata types, to stay focused on the main theme, since for the interested- such literature is aplenty available via various channels, including but not limited to, internet search.

VI. References

[1] Hobbs, Jerry R.; Walker, Donald E. and Amsler, Robert A. (1982). "Natural language access to structured text". Proceedings of the 9th conference on Computational linguistics 1. pp. 127–32

[2] Adrian, Merv (2011). "Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs", Gartner Publicationshttp://imagesrv.gartner.com/summits/docs/na/master-data-management/MDM6_Conference_Brochure_2012.pdf

[3] Belsky, Gary (2012). "Why Text Mining May Be The Next Big Thing". Time MagazineCover Articlehttp://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/

[4] NACTEM (2013). National Center for Text Mining. http://www.nactem.ac.uk

[5] Manyika, James et al. (2014). "Big Data: The Next Frontier for Innovation, Competition, andProductivity". McKinsey Global Institute report, p.1 https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[6] Grimes, Seth (2013). "Big Data: Avoid 'Wanna V' Confusion". Information Week. https://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077

[7] Analytics Magazine (2014). "Advanced Analytics Fastest Growing Segment of Worldwide BI, Analytics Software Market".

[8] SAS Institute (2013). "Five Big Data Challenges". Article number 106263_S106008.031

[9] Marcus, Gary and Davisparil, Ernest (2014). "Eight (No, Nine!) Problems With Big Data", The New York Times. https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

[10] Harford, Tim (2014). "Big Data: Are We Making a Big Mistake?" The Financial Times, London. https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0

[11] Lazer, David; Kennedy, Ryan; King, Gary and Vespignani, Alessandro (2014). "The Parable of Google Flu: Traps in Big Data Analysis". Science Vol. 343

[12] K.N.C. (2014). "The Backlash Against Big Data". The Economist Explainshttps://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10

[13] Fung, Kaiser (2014). "Google Flu Trends' Failure Shows Good Data > Big Data". Harvard Business Review Blog Network.

[14] Witten, I.H. (2005) "Text mining." in *Practical handbook of internet computing*, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida

[15] Inmon, Bill (2013), "Data, Metadata and Big Data". B-eye-network. http://www.b-eye-network.com/view/16943

[16] Olavsrud, Thor (2014). "Data Scientists Frustrated by Data Variety, Find Hadoop Limiting". CIO Magazine.https://www.cio.com/article/2449814/big-data/data-scientists-frustrated-by-data-variety-find-hadoop-limiting.html

[17] IBM (2014). "The MDM Advantage- Creating Insights From Big Data". IBM Research, article number: IMM14124-USEN-01

[18] Prevosto, Virginia and Marotta, Peter(2013). "Does Big Data Need Bigger Data Quality and Data Management?". Verisk Analytics, Inc.. https://www.verisk.com/verisk-review/archived-articles/does-big-data-need-bigger-data-quality-and-data-management/

[19] Dumbill, Edd (2013). "Big Data Variety Means That Metadata Matters". Data Driven, Forbes Magazine. https://www.forbes.com/sites/edddumbill/2013/12/31/big-data-variety-means-that-metadata-matters/

[20] Manyika, James et al. (2014). "Big data: The Next Frontier for Innovation, Competition, andProductivity". McKinsey Global Institute report, p122. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[21] Oxford (2017). Definition of "Context", Oxford Dictionary Online, accessed at: http://www.oxforddictionaries.com/definition/english/context

[22] MIT Geoweb GIS (2017). Metadata sample derived from MIT Geoweb-GIS (Geographic Information Systems) service: 'Metadata Reference Information. https://arrowsmith.mit.edu/mitogp/layer/Tufts.USAPartMat06/

[23] **NISO (2017). Metadata sample derived from National Information Standards Organization (NISO) "Understanding Metadata", Dublin Core (page 3) accessed at: www.niso.org/publications/press/UnderstandingMetadata.pdf**