DBKDA 2012, March 02, 2012 – St. Gilles, Réunion

Panel discussion

On the Quality of Non-structured Data

Moderator:

Fritz Laux, Reutlingen University, Germany

Panelists:

Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany

Chris Ireland, Open University, UK

Eloy Gonzales, National Institute of Information and Communications Technology, Japan

Maria Del Pilar Angeles, Universidad Nacional Autónoma de México, Mexico

Dumitru Dan Burdescu, University of Craiova, Romania

What is Quality of Data?

Informal Definition:

If the data is useful for its intended purpose [Juran, 1989]

Hence, the Encyclopedia Britannica is of low quality for me if I want to build a DBMS?

Formal Criteria: (but no formal definition)

List of quality attributes, e.g. IAIDQ, DGIQ adopted 15 dimensions of attributes from [Wang/Strong, 1996] within 4 categories:

- accessability
- interpretation
- relevance

accuracy

NB: in German language the term information quality is prefered

What is “non-structured” Data?

Propositions:

(1) If data is completely unstructured it is „noise“

(2) we should use „variable-structured“ instead

But, valuable information might be hidden behind noise. So, there is some structure, but it might be hidden or its structure is variable over time

This makes quality measures difficult.

The task is to find invariants and

extract the structure in order to measure the quality of the data.

What are quality measures (QM) of “variable-structured” data?

In priciple we have to define QM as
a function that compares the data in question to its invariants
and produces an ordinal value, the quality measure.

Let D be a set of data (documents) and
let N be a an positive number interval
The function QM_c: D → N is a quality measure for every invariant c.

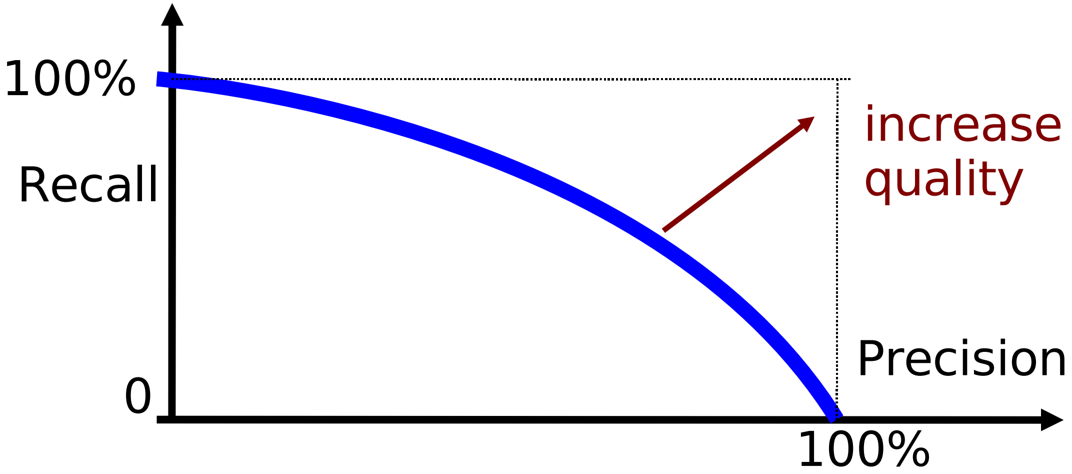
The QM function has to take in account all quality dimension
- accessability
- interpretation
- relevance
- accuracy

Example

Search quality measure: recall and precision

criteria

- ontology/thesaurus
- check actuality
- result ranking
- follow and analyse links (in case of web docs)

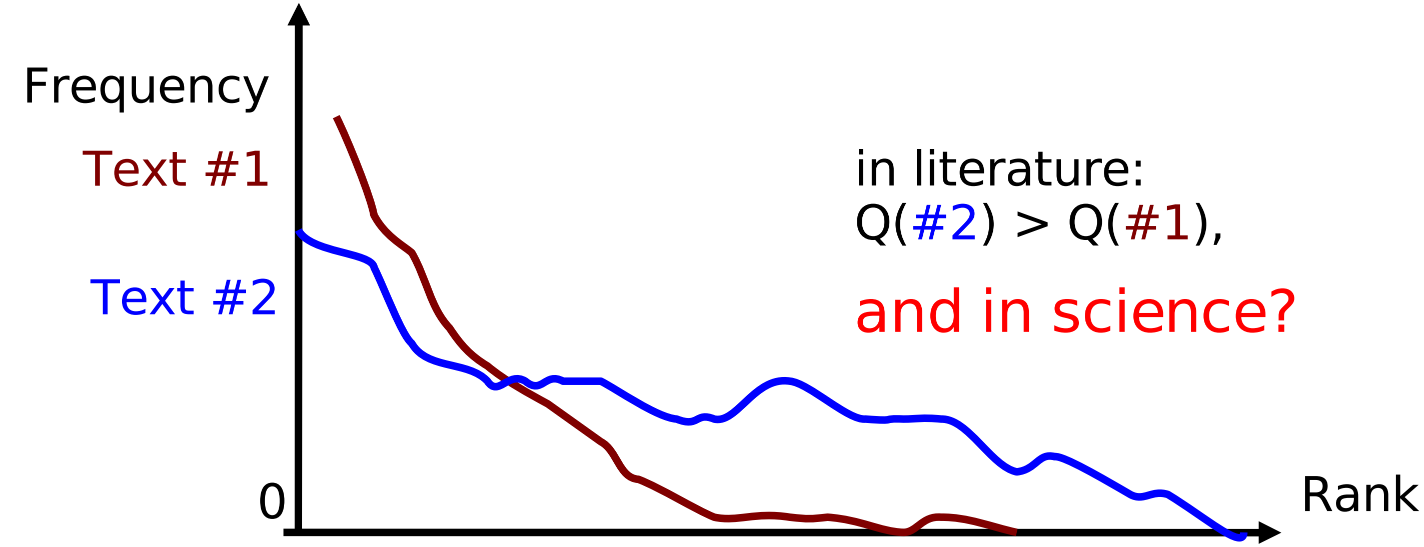


Panel Discussion: On the Quality of Non-structured Data

Example

How to assess intrinsic data quality?

Criteria
- understandability, consistency
- actuality, accuracy
- structure
- author's expertise/reputation



in literature:
Q(#2) > Q(#1),

and in science?

The Forth International Conference on Advances in Databases, Knowledge, and Data Applications
(DBKDA 2012)

February 29 - March 5, 2012 - Saint Gilles, Reunion Island

Panel: On the Quality of Non-structured Data

- status, vision and challenges -

Andreas Schmidt

Institute for Applied Computer Science
Karlsruhe Institute of Technologie
PO-box 3640
76021 Karlsruhe
Germany

Department of Informatics and Business Information Systems
University of Applied Sciences Karlsruhe
Moltkestraße 30
76133 Karlsruhe
Germany

Classification of the problem

Data quality (DQ)

- well understood for structured data
- different models, methodolgies and tools
- DQ Dimensions (amongst others):
  - Accuracy
  - Consistency
  - Completeness
  - Currency

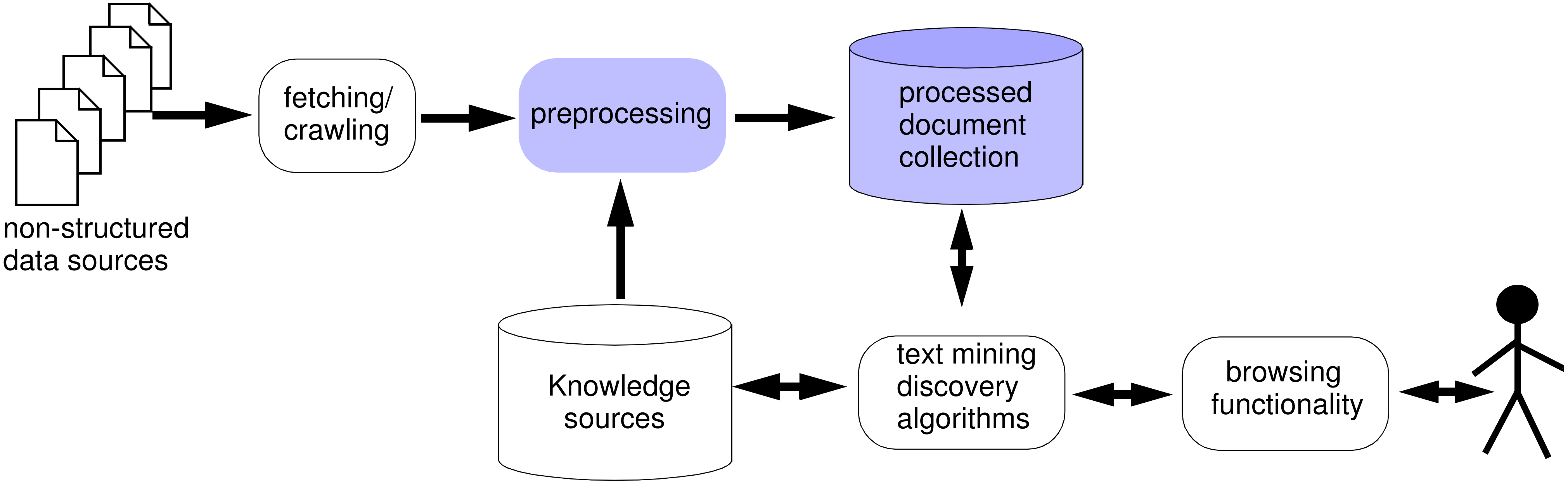
=>mostly applicable for relational data

Non-structured data

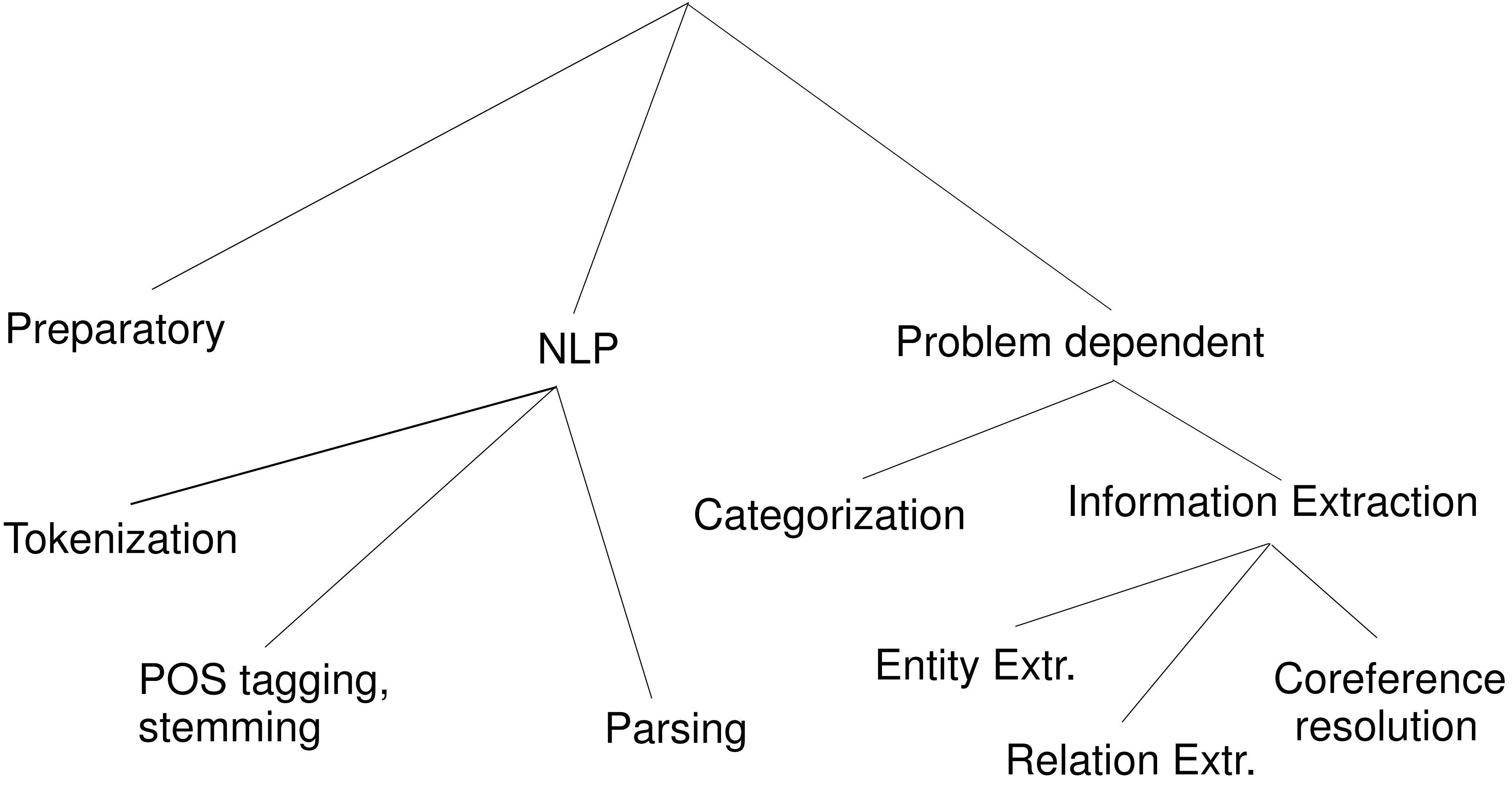
- Data with no predefined data model
- 80% of available data is unstructured
  - Email
  - Documents (pdf, word, CMS, ...)
  - WWW (blogs, news)
  - Pictures, Video, Speech recording
  - ...
- How can the quality of this data be measured ?
  => what should be measured ?

Text Mining

- Approach: Transforming the raw, unstructured data into structured data, which than can be queried or measured
- Combination of techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management.


non-structured data sources
fetching/ crawling
preprocessing
processed document collection
Knowledge sources
text mining discovery algorithms
browsing functionality

Preprocessing Tasks



What determines the quality of non-structured data?

- Technologies used to extract the data from non-structured sources
- Used methodology to assure Data Quality

non-structured data quality

=

extraction of relevant features + adequate measurement methodologies

Challenges/Open Points

- Unstructured data are one of the biggest challenges in data quality management
- Establishment of more data quality methodologies/research for non-structured data
- Adaption of such methodologies to specific domains (financial-, public sector, ...)
- Number of experimental studies to non-structured data quality

References

- Carlo Batini, Monica Scannapieco. Data Quality: Concepts, Methodologies and Techniques. Series: Data-Centric Systems and Applications, Springer, 2006
- Ronen Feldman, James Sanger: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured data, Cambridge Press, 2007
- Carlo Batini, Cinzia Cappiello, Chiara Francalanci, Andrea Maurino: Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, Vol. 41, No. 3, Article 16, Publication date: July 2009.
- Andrew McCallum: Information extraction - Distilling Structured Data fom Unstructured Text, ACM Queue, November 2005.
- Carlo Batini, Daniele Barone, Federico Cabitza, Simone Grega. A Data Quality Methodology for Heterogenous Data. International Journal of Database Management Systems, Vol. 3, No.1, February 2011

Backup slides

Example: IE - template element task

(from Feldman & Sanger 2007, pp. 98)

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. L.J.G. is headquartered in the Maddox family's hometown of La Jolla, CA.

```
entity{
        ID = 1
        NAME = "Fletcher Maddox"
        DESCRIPTOR = "Former Dean of
                     UCSD Business School"
        CATEGORY = person
}
entity{
        ID = 2
        NAME = "La Jolla Genomatics"
        ALIAS = "LJG"
        DESCRIPTOR = ""
        CATEGORY = organization
}
entity{
I       D = 3
        NAME= "La Jolla"
        DESCRIPTOR = "the Maddox family hometown"
        CATEGORY = location
}
```

IE - template relationship task (TR)

employee_of(Fletcher Maddox, UCSD Business School)

employee_of(Fletcher Maddox, La Jolla Genomatics)

product_of(Geninfo, La Jolla Genomatics)

location_of(La Jolla, La Jolla Genomatics)

location_of(CA, La Jolla Genomatics)

Example: IE - template element task

(from Feldman & Sanger 2007, pp. 98)

***Original Sentence*:** Mr. Eskew was Vice President of Worldwide Sales for Sandpiper Networks, which was recently acquired by Digital Island where he created the worldwide sales strategy.

***After Part of Speech Tagging*:** <Prop>Mr. Eskew</Prop><Verb>was</Verb><Prop>Vice President </Prop><Prep>of</Prep><Prop>Worldwide Sales</Prop> <Prep>for</Prep><Prop>Sandpiper Networks </Prop> which <Verb>was</Verb><Adv>recently</Adv><Verb>acquired</Verb> <Prep>by </Prep><Prop>Digital Island</Prop> where <Pron>he</Pron> <Verb>created</Verb><Det>the </Det><Adj>worldwide</Adj> <Nn>sales strategy.</Nn>

***After Shallow Parsing*:** NP:{Mr. Eskew} was NP:{Vice President ofWorldwide Sales} for NP:{Sandpiper Networks} which was ADV:{recently} V:{acquired} by NP:{Digital Island} where NP:{he} V:{created} NP:{the worldwide sales strategy}

***After Named Entity Recognition*:** Person:{Mr. Eskew} was Position:{Vice President of Worldwide Sales} for Company:{Sandpiper Networks} which was ADV:{recently} V:{acquired} by Company:{Digital Island} where Person:{he} V:{created} NP:{the worldwide sales strategy}

***After Merging (Coreference Resolution)*:** Person:{Mr. Eskew} was Position:{Vice President of Worldwide Sales} for Company:{Sandpiper Networks} which was ADV:{recently} V:{acquired} by Company:{Digital Island} where Person:{Mr. Eskew} V:{created} NP:{the worldwide sales strategy}

***Frames Extracted*:**

Frame Type: **Acquisition**
Acquiring Company: Digital Island
Acquired Company: Sandpiper Networks
Acquisition Status: Historic

FrameType: **PersonPositionCompany**
Person: Mr. Eskew
Position: Vice President of Worldwide Sales
Company: Sandpiper Networks
Status: Past

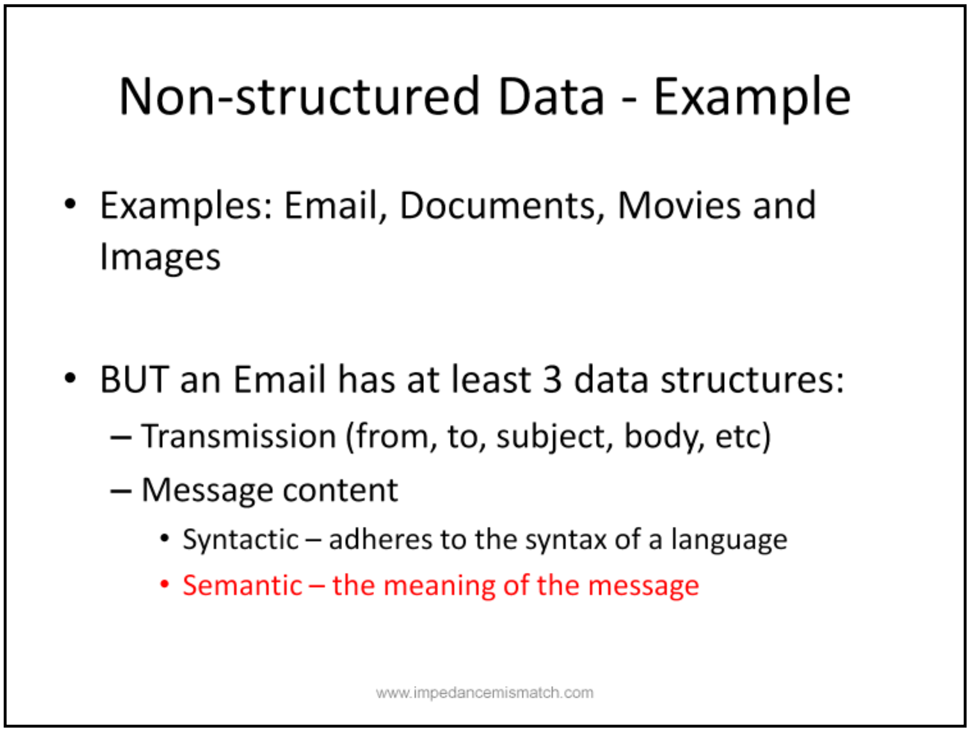
On the Quality of Non-structured Data

Chris Ireland
ImpedanceMismatch.com

www.impedancemismatch.com

Spend a few minutes exploring why this is such a difficult task…

It’s a Philosophical quagmire!



Non-structured Data - Example

- Examples: Email, Documents, Movies and Images
- BUT an Email has at least 3 data structures:
  - Transmission (from, to, subject, body, etc)
  - Message content
    - Syntactic – adheres to the syntax of a language
    - Semantic – the meaning of the message

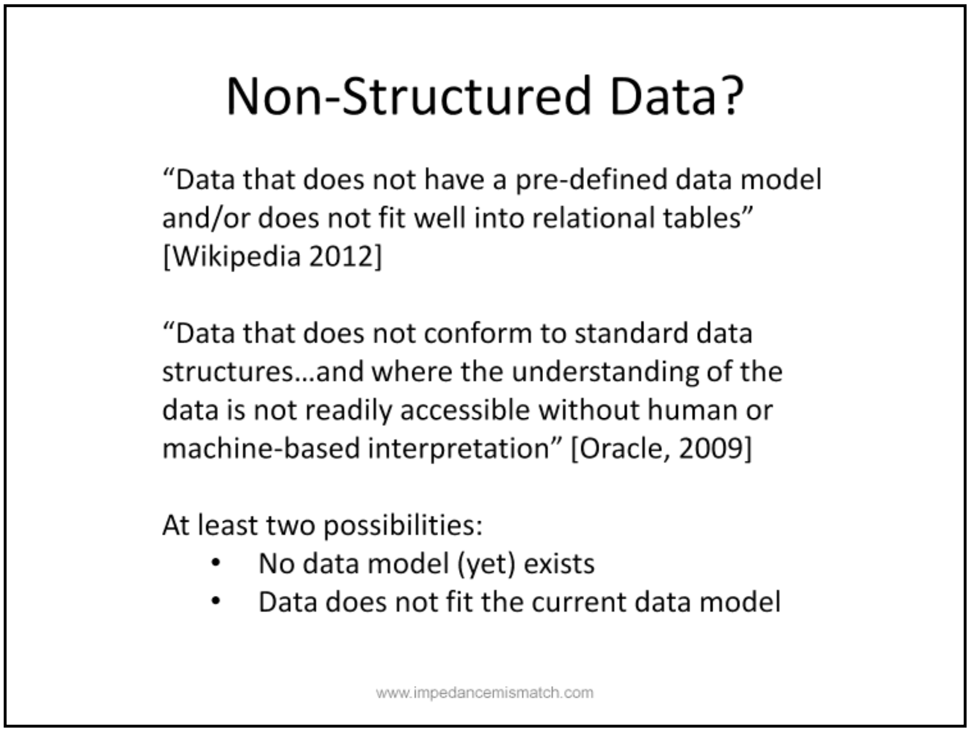
www.impedancemismatch.com

3 Different Abstractions of an Email

Each of the three data structures is important in a particular context:

- Transmission system
- Spelling and grammar checking
- Conveying a message

So if it's in a computer it has a structure – when we talk about non-structured data perhaps we are referring to the message body?



Non-Structured Data?

“Data that does not have a pre-defined data model and/or does not fit well into relational tables” [Wikipedia 2012]

“Data that does not conform to standard data structures…and where the understanding of the data is not readily accessible without human or machine-based interpretation” [Oracle, 2009]

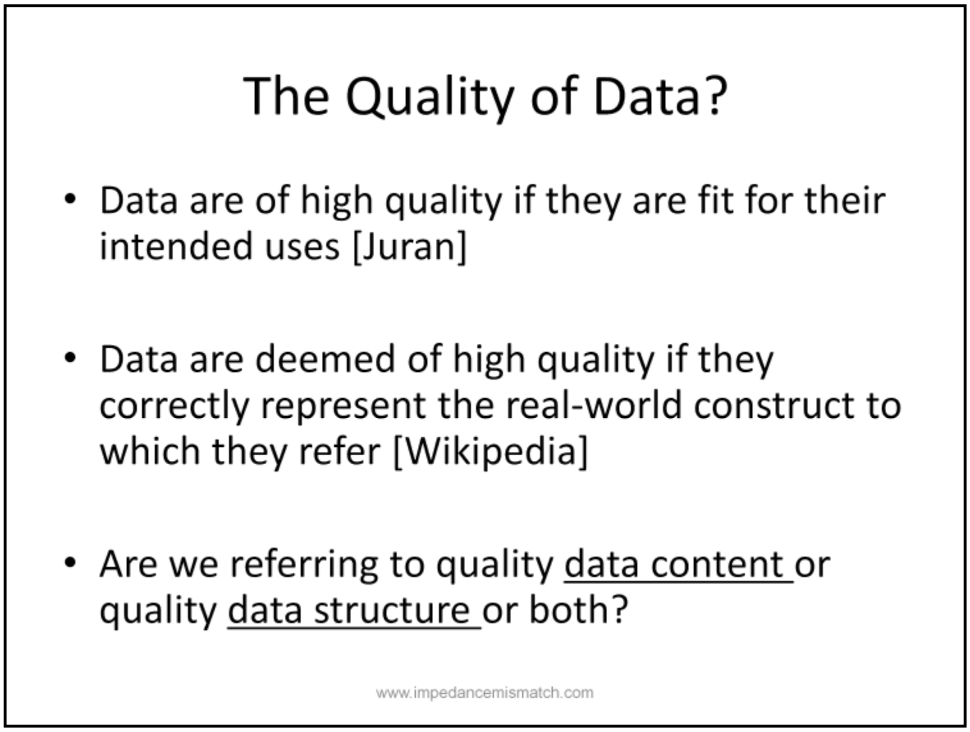
At least two possibilities:

- No data model (yet) exists
- Data does not fit the current data model

There are at least two possibilities, non structured data is data that:

1.Does not (yet?) have a data model – in other words is beyond our current understanding
2.Does not fit an existing data model – in other words our current understanding is not comprehensive

In both cases the solution is to improve our current understanding and produce or improve a data model.
Data will then cease to be non-structured! <- This is how we cope with uncertainty



The Quality of Data?

- Data are of high quality if they are fit for their intended uses [Juran]
- Data are deemed of high quality if they correctly represent the real-world construct to which they refer [Wikipedia]
- Are we referring to quality data content or quality data structure or both?

www.impedancemismatch.com

To what are we referring when we use the term quality data?

Quality content?

Data content is relevant/reliable/etc

But it is possible to have quality data in an unstructured form

e.g. pages of a book stored out of sequence

So the structure of a book is preserved but not the structure of a story!

Quality structure?

The data fits some pre-defined ontology/model

But it is possible to have poor quality data in a structured form

e.g. partially complete or incorrect data values

(that are still valid in some sense e.g. a wrong – but valid- post code)

Remember in the old days when we used to tune a radio! Hidden in all that noise might have been a signal e.g. Morse code. If you were listening for that signal then you might hear it (because you were expecting it i.e. you have a data structure for assessing the quality of a signal), otherwise you would just hear scrambled noise. In other words one persons noise is another persons high quality data!

Attempted definition of data quality (with problems):

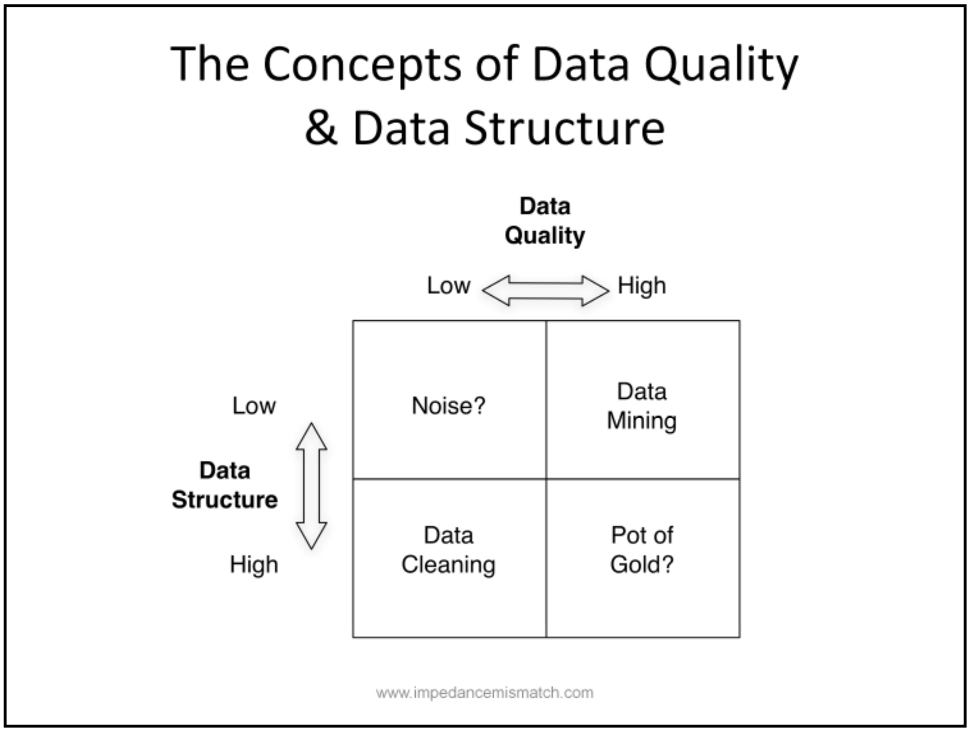
High quality data – data that fits our data model?

But we can have high quality data that does not fit our model

Low quality data – data that does not fit our data model?

But we can have low quality data that fits our data model

Consequently the concepts of structure and quality are independent



The Concepts of Data Quality & Data Structure

Data Quality

Low ⟷ High

Data Structure	Low	High
Low	Noise?	Data Mining
High	Data Cleaning	Pot of Gold?

www.impedancemismatch.com

The concepts of quality and structure are INDEPENDENT:

Approaches to the quality of non-structured data can be descriptive or prescriptive

Descriptive – understand the structure of data

Data with no discernable data structure is more difficult to assess because we have no baseline for assessment of such things as its meaning

i.e making sense of it

Techniques include Data mining, analytics, etc.

Prescriptive – fit data into a pre-defined structure

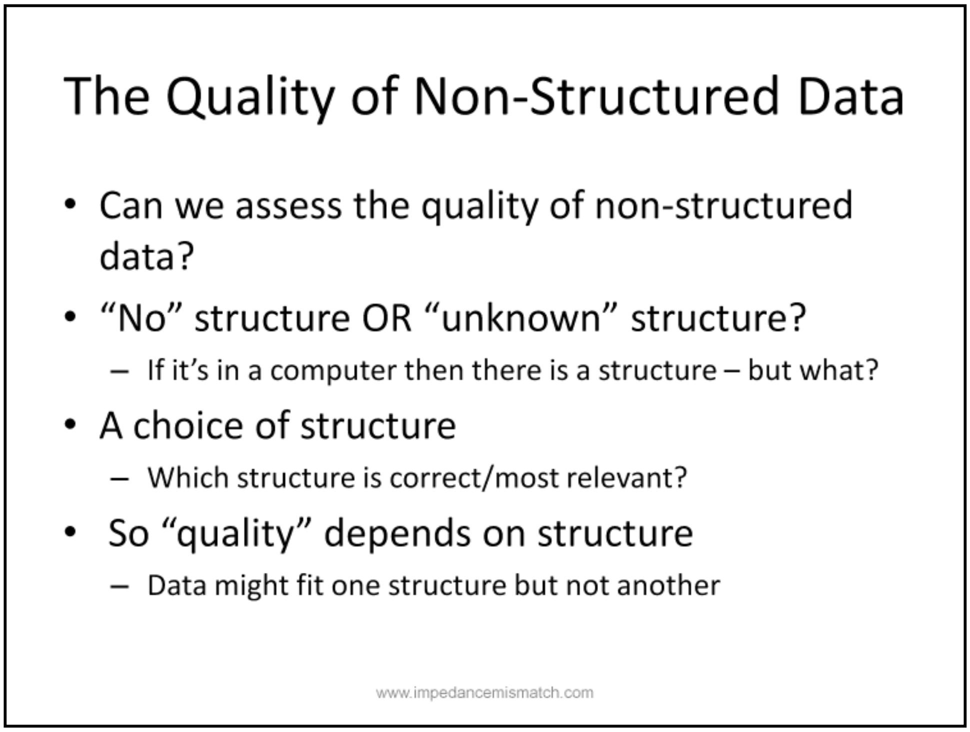
Non structured data content can be interpreted in terms of an existing data structure

i.e. to what extent does a data stream adhere to a particular data structure?

Techniques include Data Cleaning

PROBLEM: How do we know that we are using an appropriate data structure?

i.e. is the meaning we attribute to the data the correct meaning?



The Quality of Non-Structured Data

- Can we assess the quality of non-structured data?
- "No" structure OR "unknown" structure?
  – If it's in a computer then there is a structure – but what?
- A choice of structure
  – Which structure is correct/most relevant?
- So "quality" depends on structure
  – Data might fit one structure but not another

www.impedancemismatch.com

Questions to think about…

Does non-structured data actually exist?
The problem is not that the data is unstructured
we just don't yet have a structure for interpreting the data
(Does the term "data" imply a structure i.e. what separates data from noise?)

Is it possible for non-structured data to be stored in a computer?
We can talk about data with a poor or unknown structure but that it not the same as NO structure.

Does it make sense to talk about the quality of unstructured data?
if we do not have a structure to understand that data?

If we have high quality data but we do not understand it then how do we know that it is of high quality?
We might assume its relevant because it is timely and that it is timely because we are looking for it
e.g. a beep on a wireless transmission may be assumed to be a Morse signal
BUT it may also be random noise.
BUT without a structure we do not know what to listen for and when to expect it.
However our confidence in a model is improved when data fits the model

So should the question be about a quality data structure?
If we have a poorly defined data structure then if some data is of high quality in relation to that structure
SO WHAT! We have high quality, poorly defined data! (The best worst thing!)

How do we know a data structure is of high quality?
If we have a well defined data structure we can explore the quality of data in relation to that structure

So a data structure is important for understanding the quality of data in a meaningful way
Consequently I question whether it is beneficial to talk about the quality of non-structured data.
Rather perhaps we should ask: **what is the quality of our understanding?**

On the Quality of Non-structured Data

Dumitru Dan Burdescu,
Marian Cristian Mihăescu,
Computer Science and Information Technology Department,
University of Craiova, Romania

General Domains Where Data Quality Is Important

- **Artificial Intelligence & Intelligent Systems**
  - Search Strategies
  - Knowledge Based Systems
  - Machine Learning and Neural Networks
  - Agents
  - A.I. Planning
- **Information Management (Databases)**
  - Information Retrieval
  - Semantic Web

General Research Issues

- Data gathering techniques and technologies
- **Data quality**
- Data modeling
- Knowledge representation
- Data/Results Visualization and Interpretation
- Continuous model improvements

Data Quality

- **Poor quality data:**
  - implies huge money losses
  - is a leading cause of IT project failure or underperformance
- **Solutions:**
  - Have a collection of tools, techniques, and processes which prepare data for use.
  - Use application custom designed procedure for quality assessment of data
  - Use quality assessment metrics

Data Quality

- **Solutions**
  - **Standardization** from the point of view of data representation;
  - **Verification** from the validation point of view
  - **Enriching** - This means recognizing a document contains verification and/or standardization characteristics.

Current Trends

- **Linked Data** –
  - connect related data that wasn't previously linked
  - lower the barriers to linking data currently linked using other methods
  - describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web

Current Trends

- **Open Data**
  - Develop open-source solutions for promoting the use of statistical data.
  - Improve data and metadata accessibility and overall quality in support of research, policy making, and transparency in may fields: economics, finance, healthcare, education,
  - Integrate data from many sources

Conclusions

- The following activities need to be performed on data such that the quality increases:
  - **Discover** the existence of data
  - **Access** the data for research and analysis
  - **Find** detailed information describing the data and its production processes
  - **Access** the data sources and collection instruments from which and with which the data was collected, compiled, and aggregated
  - **Effectively communicate** with the agencies involved in the production, storage, distribution of the data
  - **Share** knowledge with other users

National Institute of Information and Communications Technology

研究開発推進ファンド
連携プロジェクト（タイプⅡ）

Information Asset Management for data citation among large-scale, multi-domain and heterogeneous databases

Information Services Platform Laboratory

Panel Discussion: On the Quality of Non-structured Data

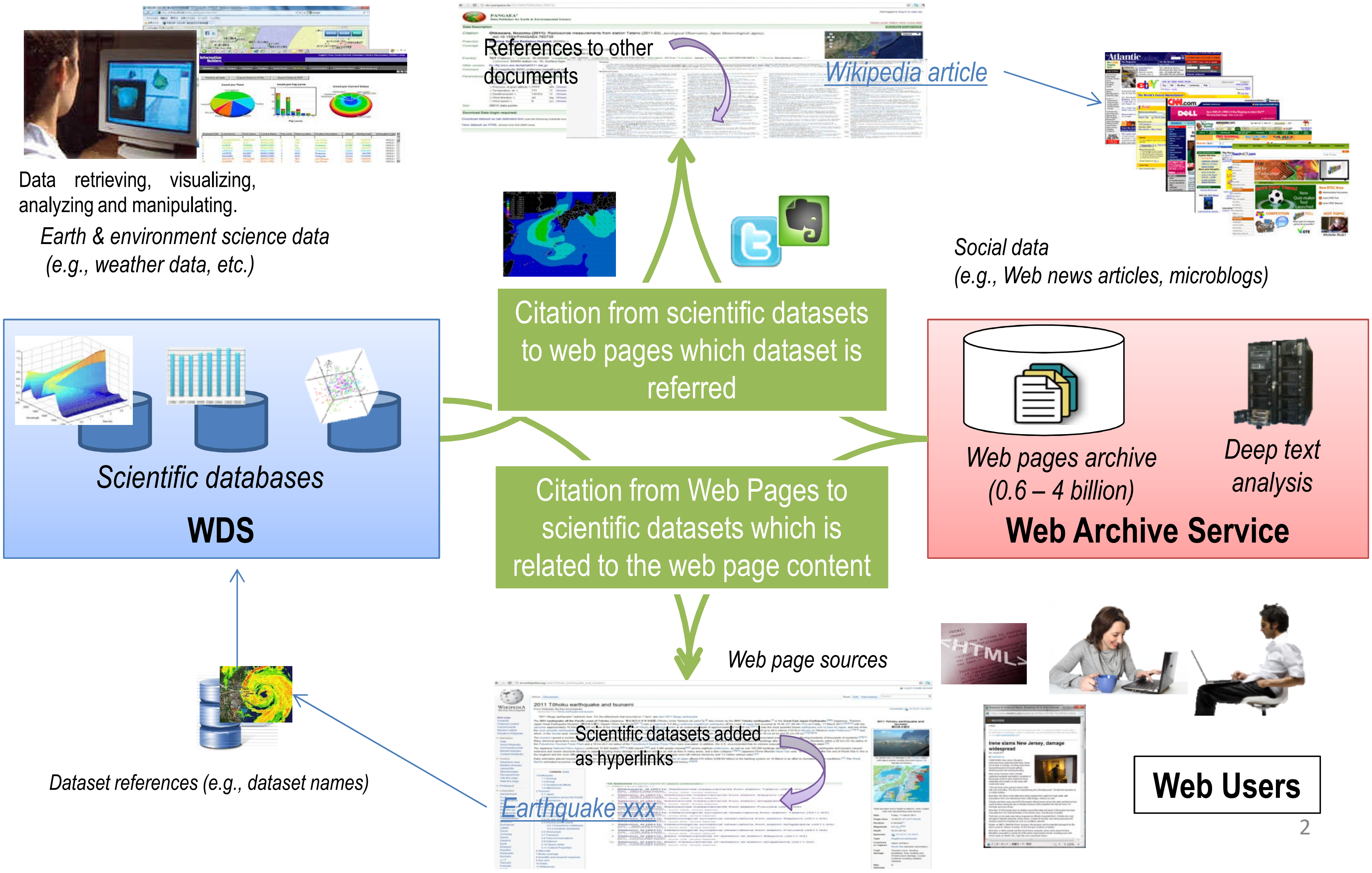
DBKDA 2012, Saint Gilles, Reunion Island

March 1st ,2012

Data Citation

Generate links and references between web documents.

Purpose: Credit and accountability for data authors, Aids reproducibility of science, i.e. direct, unambiguous connection to the precise data used.



References to other documents

Wikipedia article

Data retrieving, visualizing, analyzing and manipulating.

Earth & environment science data (e.g., weather data, etc.)

Social data (e.g., Web news articles, microblogs)

Citation from scientific datasets to web pages which dataset is referred

Citation from Web Pages to scientific datasets which is related to the web page content

Scientific databases

WDS

Web pages archive (0.6 – 4 billion)

Deep text analysis

Web Archive Service

Web page sources

Scientific datasets added as hyperlinks

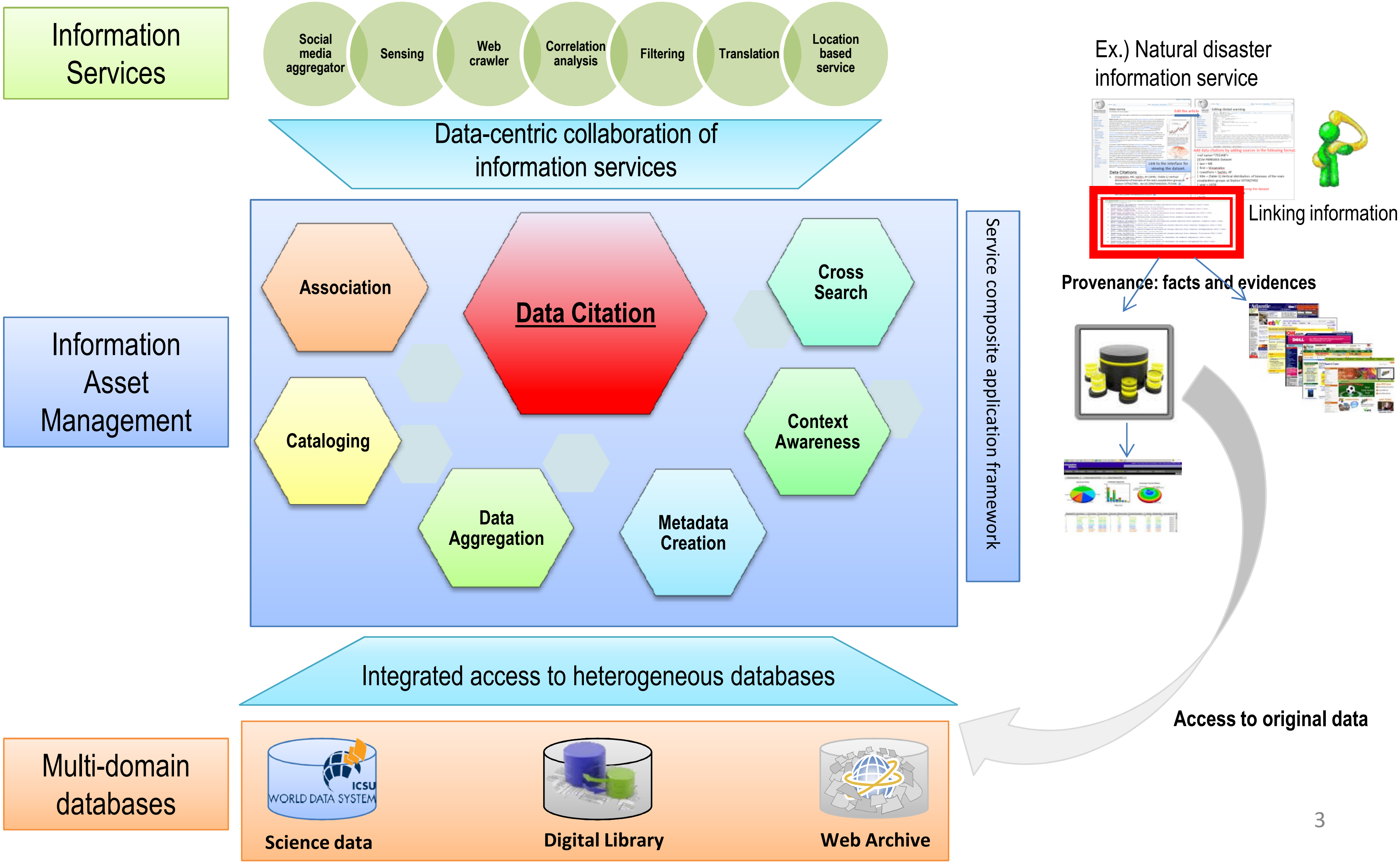
Earthquake xxx

Dataset references (e.g., dataset names)

Web Users

Information Services Platform Laboratory

A service computing platform for building and utilizing information assets.
Data Citation, as an information asset, has become critical for improving understanding and reliability.



Information Services

Social media aggregator

Sensing

Web crawler

Correlation analysis

Filtering

Translation

Location based service

Data-centric collaboration of information services

Information Asset Management

Association

Data Citation

Cross Search

Cataloging

Context Awareness

Data Aggregation

Metadata Creation

Service composite application framework

Integrated access to heterogeneous databases

Multi-domain databases

Science data

Digital Library

Web Archive

Ex.) Natural disaster information service

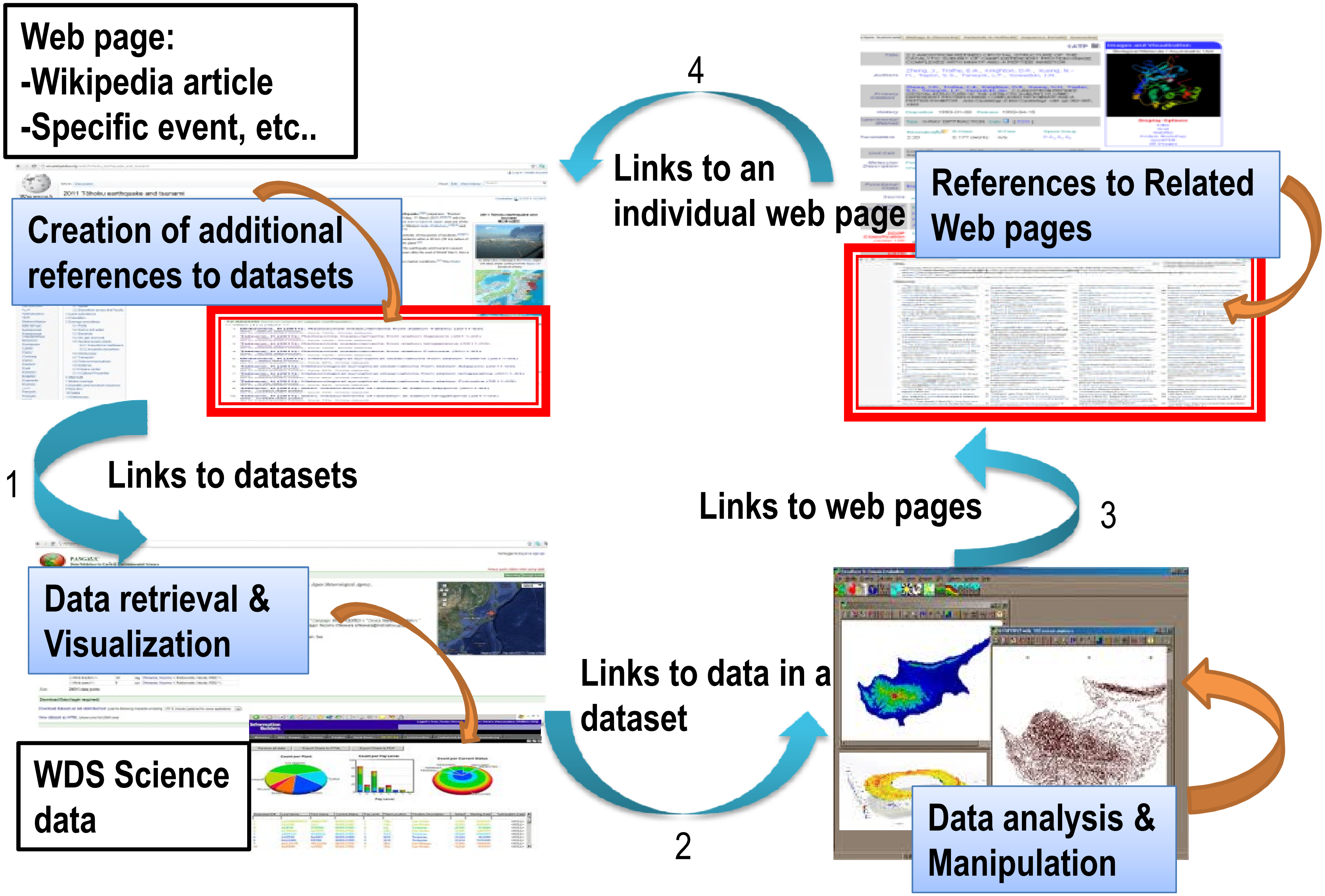
Linking information

Provenance: facts and evidences

Access to original data

Vision: Technology and Application

Data Citation



Web page:
-Wikipedia article
-Specific event, etc..

Creation of additional references to datasets

1 Links to datasets

Data retrieval & Visualization

WDS Science data

Links to data in a dataset 2

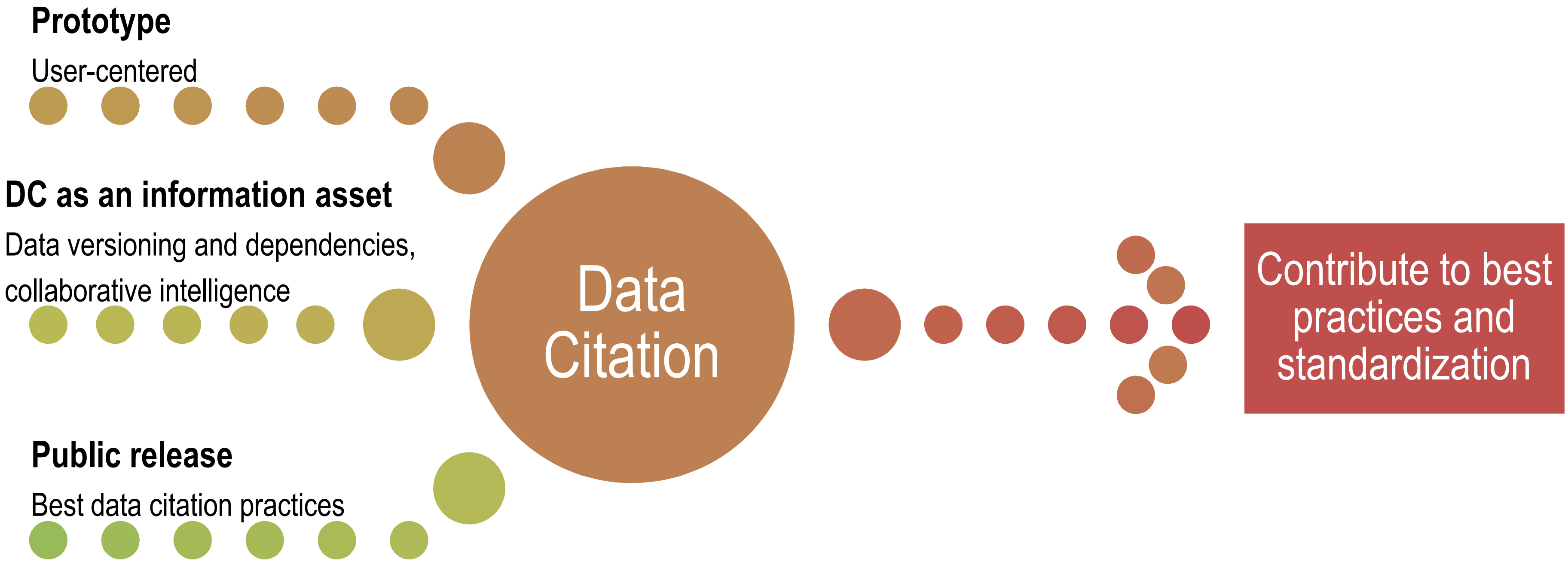
Data analysis & Manipulation

Links to web pages 3

References to Related Web pages

4 Links to an individual web page

Objective



Prototype

User-centered

DC as an information asset

Data versioning and dependencies, collaborative intelligence

Public release

Best data citation practices

Data Citation

Contribute to best practices and standardization

Challenges

- Complexity of data management
  - Continuous data, subsets, on-demand data
- Complexity of data networks
- Best practice guide for data citation
- Establishing and sustainability of ‘Data Citation Service’
- Socio-cultural challenges
- Archival of dataset cited by user
- Performance and Scalability

On the Quality of Non-Structured Data

Moderator:
Friedrich Laux, Reutlingen University, Germany

Guest panelists:
Gledson Elias, Federal University of Paraíba, Brazil
Maria Del Pilar Angeles, Universidad Nacional Autonama de Mexico
Eloy Gonzales, National Institute of Information and Communications Technology, Japan
Alain Pirotte, Universite de Louvain, Belgium
Todd Eavis, Concordia University, Canada
Andreas Schmidt, Karlsruhe University of Applied Sciences, Germany

- Over 80% of corporate information is non structured and remains out of reach for current database management and data quality technology.

- There is a tendency to believe that unstructured data as it is no under specific data model constraints or concurrency/consistency control (acid properties) it may lack of quality…

- ….is it true?

If information in unstructured document is presented in a logical pattern, humans can ‘read’ and ‘understand’ it, but current databases, BI and reporting technology cannot…

There are different approaches to take advantage of unstructured data within organizations. Two of them will be briefly explained as follows:

Identify a Domain expert

Identify useful data and categorized.

Extract unstructured data into a structured form.

Data cleansing and management for their integration with structured data sources.

- OR
  - Identify a Domain expert for categorization of useful data
  - The use of a parser to identify, evaluate , and find the best pattern.
  - The resulting XML package is then transformed to create the final, persistent XML representation of the document metadata.
  - Metadata normalization, again through pattern matching.
  - The resulting XML package is then transformed to create the final, persistent XML representation of the document metadata.

- However….is the process too complex??…
  - Format conversions are required depending on the source documents and the needs of the system
  - Some search applications require additional linguistic and token processing, beyond that provided by the search engine itself.
  - Care should be taken when merging documents from disparate collections into a single search engine experience.
- DATA QUALITY DEPENDS ON THE DOMAIN AND EXPERTISE, which is good for an expert might be poor for other

- **At the end of the process will it increase de quality of Unstructered data?**

- **Handling Data Quality in Unstructured Data**

- The patterns are too complex to ‘read’ even for today’s data management software.

- http://zorallabs.com/services/data-quality-management-of-unstructured-data

- http://www.searchtechnologies.com/document-processing-best-practices.html