DATA-CENTRIC VIEW IN E-SCIENCE INFORMATION SYSTEMS

***Gregor Erbach*[1*]**

[*1]*German Research Center for Artificial Intelligence, Language Technology Lab*
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
Email: gregor.erbach@gmail.com

ABSTRACT

Network approaches in Current Research Information Systems support the shift from a document-centric to a data-centric view, which acknowledges the primacy of data in the scientific process. e-Science holds the promise of a complete, data-centred documentation of the scientific process.

Keywords: Research information systems, Networks, e-Science

1 INTRODUCTION

1.1 Current Research Information Systems

Current Research Information Systems (CRIS) contain information about various aspects of research activities, notably research actors (persons, organisations, funding agencies), research activities (projects), tools (software systems, instruments, infrastructures), resources (data, knowledge bases), concepts (theories, analytic categories, technologies), and outputs (publications, prototypes, data) and the relations between these entities. Examples of such relations are:

- works_for (person, organisation)
- author_of (person, publication)
- owner_of (organisation, dataset)

Such information is commonly represented in relational databases (entity-relationship model), object-oriented databases, or formal ontologies with relational attributes. Many European countries have adopted the Common European Research Information Format (CERIF) as a standard for their current research information systems (Assersen, Jeffery, & Lopatenko, 2002) (Jeffery, Asserson, & Lopatenko, 2002).

1.2 Views of Research Information

The entities and relationships constitute a large network, through entities that are either directly or indirectly connected. The network view enables the application of network analysis techniques for the discovery of derived relationships. The traditional network analysis method in library science is *bibliometrics*, which studies authorship (author/publication) and citation (publication/publication) relationships to determine derived relations such as co-citation or impact factors. This approach will be referred to as the *document-centric view* of research information. Similar methods have been applied to the network constituted by hyperlinks on the WWW and form the basis of relevance ranking algorithms for search engines. Social network analysis studies the relationships between people (Wasserman & Faust, 1994), and there has been much research recently on the properties (scale-free, small-world) of such networks (Barbarasi, 2002). In the context of research information, network analysis can be applied to determine relationships among people, based on their joint projects, organisations they work for, or data sets they work with. This is a person-centric view. Likewise, it can be used to analyse relationships among data sets based on the people who work with them, the publications that reference them, and the data analysis methods applied to them. This will be referred to as a *data-centric view of research information.*

2 THE PRIMACY OF DATA

It is important to observe that the data-centric view is not just any view of research information, but that it enjoys special status because of the primacy of data in scientific research. After all, the analysis, interpretation, modelling, and understanding of sets of observable data are what the scientific endeavour is all about.
Scientific data sets can take many different forms for different fields of research. Examples are:

- data streams from scientific instruments such as telescopes, seismographs, Earth observation satellites, accelerators, and hadron colliders
- measurements of physiological functions, such as FMRI brain scans and eye movements
- geo-data
- socio-economic data, such as GDP, stock market and currency market data and indices, employment statistics, and survey data
- human interaction data , such as video/audio recordings
- text and audio-visual data , such as corpora

Some data occur naturally (weather, astronomical events); others are the incidental results of human activity (socio-economic data), and yet others are produced in the interest of scientific investigation (experimental data). Whatever the origin of the data, it must be documented through meta-data with an explanation of when, where, and how they were captured, with which instruments, by whom, and how the data is coded. Such meta-data are applicable to entire data sets as well as individual data points (time and location stamps). Data are often not captured and represented in a theory-neutral way, but the *coding of the data* is based on certain theoretical assumptions and the interests of the researchers.

2.1 Data Analysis Produces More Data

Much of scientific activity involves the analysis, aggregation, and interpretation of data, and in the process, the generation of new data sets. These new data sets arise either from the application of formal algorithms (e.g. statistical analysis packages or visualisation tools) to the data or by the mental activity of the researcher who tries to make sense of the data by grouping them into classes and linking them to hypotheses and theories. We refer to the former as *transformation* of the data and to the latter as *annotation* of the data. In either case, it is desirable that such transformed data sets or annotations exhibit proper meta-data indicating on which data set they depend.

There is of course the possibility for multiple annotations of the same data and for incremental annotations of annotations. According to this view, a scientific publication is the ultimate form of annotation, linking data sets with interpretation and previously published theories. Note that traditionally, the data sets used are not formally referenced in publications but are referenced through informal descriptions, in a document-centric way, through a reference to the publication in which the data set is described, or more recently, by providing a URL where the data set can be downloaded.

3 E-SCIENCE AND THE SCIENTIFIC METHOD

The scientific method demands that research results be documented in such a way that they are reproducible by other researchers. This implies that:

- data sets used must be accessible
- data sets must be documented with meta-data
- data-sets used must be properly referenced in a publication
- data analysis methods must be documented
- data analysis methods must be accessible
-

Current activities in e-science support such requirements by moving scientific activity to the web or the GRID. E-science infrastructures enable data to be captured in one place, stored in a different place, and processed in yet another location, supported by a computing grid that provides for high-speed networking and by infrastructures that control access to and payment for data and services, and most importantly for

data science, provide a description of the data and services. Semantic Web standards have been proposed for describing data sets and grid services (Hendler, 2003).

4 RELATED WORK

Our own work has concerned the development of the information system LT-World (http://www.lt-world.org/) for the area of language technology, which uses the Ontology Web Language OWL to represent the data structures (Uszkoreit, Jörg, & Erbach, 2003) (Jörg & Uszkoreit, 2005). Protégé was used as a data modelling tool for research information in general and domain concepts. In the user interface, the relations presented as hyperlinks in user interface for easy navigation, for example, from a person to his/her projects or software systems or from a patent to the owner or inventor. The project *Project Intelligence* (Grobelnik & Mladenic, 2002) has used data analysis techniques (data mining, clustering) to identify and visualise relationships between different projects, countries, and funding programmes in the EU's Framework Research Programme.

The project *IST World* (Erbach, Grobelnik, Jermol, Jörg, & Uszkoreit, 2005) aims to implement a CERIF-based information system for European research activities in information society technologies. Data will be represented in semantic web standards and data mining and social network analysis methods applied for data analysis.

5 CONCLUSION

5.1 Impact on Publication and Documentation Practices

A view that makes data sets first class objects requires certain changes in publication and documentation practice, for example the records for projects and publications in e-science information systems should be extended with new fields "used_dataset" and "generated_dataset" and the record for datasets with a field "depends_on_dataset."

For publications whose authors did or do not properly reference the data sets they used or produced, information extraction methods can be applied to extract natural languages references to data sets such as "the CHILDES child language corpus" or "the Map Task dialogue data" from text.

5.2 Benefits of a Data-Centric View

A data-centric view helps researchers find out more information about the data sets that they are working on:

- who else has worked with it
- which data analysis methods have been applied to the data (methods, domain-concepts)
- which project has worked with the data
- are analysed/interpreted/aggregated versions of the data available
- which publication exist about the data
- which other data sets have been used by persons/projects who used my data set
- which other data sets have been cited by publications which cite my data set
- which other data sets have been treated with the analysis methods I use

Thereby it promotes interaction with other researchers, the spreading of best practice in data analysis and annotation, and interdisciplinarity.

6 REFERENCES

Asserson, A., Jeffery, K.G. & Lopatenko, A. (2002) CERIF: Past, Present and Future: An Overview. *Gaining Insight from Research Information, 6th International Conference on Current Research Information Systems,* Kassel, Germany.

Barabasi , A.-L., (2002) *Linked: How Everything Is Connected to Everything Else and What It Means.* New York: Perseus Books Group.

Erbach, G., Grobelnik, M., Jermol, M., Jörg, B. & Uszkoreit, H. (2005) Network Approaches in Current Research Information Systems. *Proceedings of eChallenges 2005,* Ljubljana, Slovenia.

Jeffery, K. G., Asserson, A. & Lopatenko, A. S. (2002) Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories. *Gaining Insight from Research Information, 6th International Conference on Current Research Information Systems,* Kassel, Germany.

Jörg, B. & Uszkoreit, H. (2005) The Ontology-based Architecture of LT World, a Comprehensive Web Information System for a Science and Technology Discipline. *Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven im europäischen Wissensmarkt. 27. Online Tagung der DGI.* Frankfurt, Germany.

Hendler, J. (2003), Science and the Semantic Web. *Science 299,* 520-521.

Grobelnik, M. & Mladenic, D. (2002) Approaching Analysis of EU IST Projects Database. *International Conference on Information and Intelligent Systems IIS-2002.* Varazdin, Croatia.

Uszkoreit, H., Jörg, B. & Erbach, G. (2003) An Ontology-based Knowledge Portal for Language Technology. *Proceedings of ENABLER/ELSNET Workshop "International Roadmap for Language Resources".* Paris, France.

Wasserman, S. & Faust, K. (1994) Social Network Analysis : Methods and Applications, *Structural Analysis in the Social Sciences Series.* Cambridge: Cambridge University Press.