

Math&Industry: The Challenge to Build a CRIS Experiences in the Fields of Applied Mathematics and Renewable Energies

Robert Roggenbuck

Institut für wissenschaftliche Information Osnabrück e.V. (IWI), Germany

Abstract

The Math&Industry project, started 2001, has the overall aim to bring the complete information of the (applied) mathematics program of the German Research and Education Ministry (BMBF) into the Web and to make it accessible and usable for applications as well as for the sciences (www.mathematik-21.de). The concept of Math&Industry covers information about the funding program as well as presentations of the projects. Semantic Web technologies are used to make the information accessible in an easy way. In 2007 we applied this approach to the research field “Renewable Energies and Efficient Usage of Energy”.

We will show the advantages and problems of our approach and evaluate the possibilities to combine our scheme with CERIF. This will allow us to outline the possible contribution of our concept to the ERA.

1. Introduction / Problem

Today, research is the key for innovation, in industry, in management and services. Up to the era of the Web, information about research was focused on research *products*, e.g. patents, technologies, publications and software (esp. for mathematics and computer sciences). The advent of the Internet and WWW has changed dramatically the information and communication capabilities. Information about research can be presented and combined in a more complete, flexible and systematic way. And the Web provides excellent opportunities for a complete and sophisticated presentation of research activities.

The sensitivity for research information has increased during the last years: Many international and national initiatives and projects are active in the field of research information systems, e.g. Scirus¹ as a commercial example, and Forschungsportal.net², Wissenschaftsportal b2i³ and FÖKAT⁴ as activities in Germany. About the CERIF model we will talk later.

1 <http://www.scirus.com>, science specific search engine, run by Elsevier

2 <http://www.forschungsportal.net>, full-text index on servers of public financed research in Germany

3 <http://www.b2i.de>, portal for library-, book- and information sciences

4 <http://oas2.ip.kp.dlr.de/foekat/foekat/foekat>, catalogue of projects funded by BMBF and BMWi (German Federal Ministry of Economics and Technology)

2. Analysis / Information and Communication in the Web: Databases, Semantic Web and Search Engines

We need a fast and structured information system which provides sufficient information for different user groups. This system must enable the projects to supply their information in a handy way. To provide an easy access to the information via the Internet, we need a central point - a web portal where all the information is gathered and visible. On the other hand we have the projects as information providers. So, the solution is a decentralized system consisting of:

- (structured) information about the program,
- (structured) information about the projects provided by the institutions carrying out the projects.

There are basically two possibilities to add semantic descriptions to the provided information: a database scheme, and Semantic Web technologies.

Databases are the probate (and historically the first) means for the management of structured data, e.g. a catalogue of a library.

At the beginning of this century, Tim Berners-Lee proclaimed the vision of the Semantic Web (Berners-Lee et al. 2001): Each information is semantically connected with other information in the Web. The technologies of the Semantic Web contain the powerful description language RDF⁵ with its RDF Schema, RDFS. The principle of RDF is very simple. The graph model of RDF is just an amount of statements of the pattern "subject - predicate - object". Because an object can be the subject of another statement (and of course it is possible to make more than one statement about one subject), the resulting "statement cloud" can become a very complex graph. In this way it is possible to make statements about any resource in a machine readable way (as content markup) in addition to the presentation in XHTML⁶ for human beings (the presentation markup).

Databases and RDF have similar facilities to structure the information and make it accessible (RDF schemata can be transformed to databases and vice versa). The technologies of the Semantic Web seem to fit much better to our scenario, because the semantics are not hidden in an invisible database scheme.

Our concept requires presentation markup as well as content markup. Both are not trivial, especially the content markup is much more complex and formal when we have the aim to cover the whole semantic content of a webpage (and not only some simple statements). That is why it is not possible to fulfil the information needs without auxiliary tools. For our special requirements we developed a special CMS (web content management system), named WebSiteMaker.

How can the decentralized information in Math&Industry become searchable and accessible? Search engines can do this. For the most users, Google is a synonym for a search engine. 2005 more than 8.000.000.000 web documents of any format were indexed by Google (Mills 2005). Google is simple and comfortable

- *for the authors*: Nothing has to be done. The authors can publish their information in any format and put it on the Web. The only requirement: The webpage has to be attainable through a link.
- *for the users*: There is a sparse user-interface to prompt the search terms. No special skills are necessary to use Google.

5 Resource Description Framework, see <http://www.w3.org/RDF/>

6 about XHTML 1.0 see <http://www.w3.org/TR/xhtml1/>

Do we still need something like Math&Industry and a special preparation of the information?

We think, yes:

1. The concept of Math&Industry evokes more information in the Web.
2. The standardized structure of the web presentation is the foundation for special services. To provide "accessibility, usability and intuitive retrieval" it is not enough to enable a web crawler to access websites with the needed information. To allow a more specific retrieval of the wanted information, we need a sophisticated semantic annotation of the data. This has to be done in a way that the information can be automatically processed.
3. Our concept combines the automatic and the manual content analysis in decentralized systems. In other words, the Math&Industry concept uses the technology of the Semantic Web and databases as well as search engine technology.

3. Model and Solution

Now we will describe the above outlined solution in more detail. At first there is the creation of data models which form the basis of the "intelligent" services for the users. The second problem is the additional coding of the information in RDF/XML. Therefore we developed software tools to do the encoding of the information in RDF/XML and XHTML. Once the data is in the Web, we need to gather and process it for the evaluation in the web portal to build up surplus value services.

3.1 The data model of a project

We started from bottom-up, in other words with the data model of a project. Since we decided to use RDF to express the semantic, starting the modelling is very easy, because we can make sentences in the "subject - predicate - object" manner. Thus we can map them nearly directly to RDF. But before we can do this, a very detailed analysis was needed to identify the things a research project (in the field of applied mathematics) consists of. As a first step, this analysis results in the following rough structure (named groups):

1. general information on the project, overview, formal project data (like project leader, funding institution, funding receiving institution, funding period, ...)
2. products and results
3. persons and organisations contributing to the project
4. detailed description of the problem to solve by this project
5. detailed description of the problems modelling
6. detailed description of the mathematical solution

The first three groups are more general for a research project, the other three groups are more mathematics related. Each of these groups (2nd level entities) covers subgroups (entities of level 3)⁷. Each subgroup has a finer structure, e.g. the description of a person covers a large set of statements. A person has a name, also at least a first name, a contact address, etc. For our aims special expertise and skills of a person are important. This can be modelled in a natural way by a graph

⁷ 'Project' is the only top entity (level 1).

(see below).

The subgroups can be assigned to eight classes:

1. description of publications
2. description of software
3. description of persons (professional home pages)
4. description of organisations (institutional home pages)
5. description of events
6. link pages (containing links to other relevant resources)
7. glossaries
8. description of other information without any standardized inner structure (e.g. "About the project")

As an example we will show in fig. 1 a RDF graph for the (minimal) description of a person.

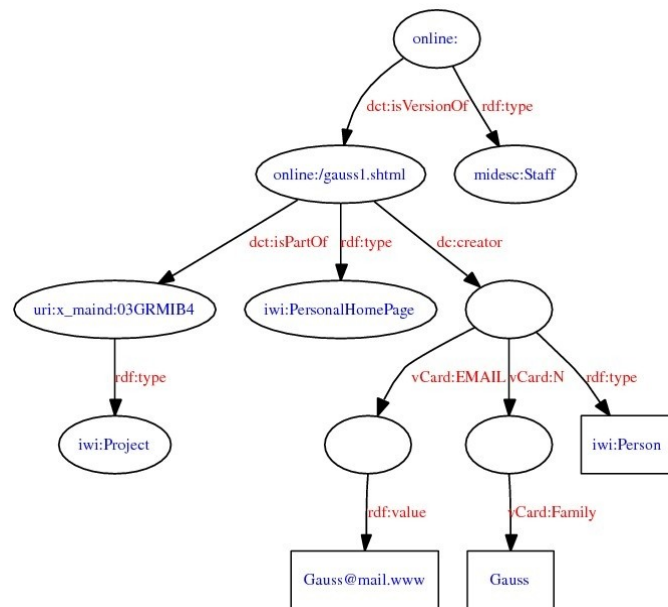


Fig. 1: RDF graph of the description of a person (part)

As mentioned above, for the web presentation we used the XML-serialisation of RDF. So, the graph in fig. 1 can be encoded in RDF/XML as shown in fig. 2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- created by MIPMPers version 2.3.1 -->
<!DOCTYPE rdf:RDF [
  <ENTITY iwi 'http://www.iwi-iuk.org/material/RDF/Schema/Class/iwi#'>
]>
<rdf:RDF xmlns="http://www.w3.org/1999/xhtml#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dct="http://purl.org/dc/terms/"
```

```

xmlns:iwi="&iwi;"
xmlns:midesc="http://www.iwi-iuk.org/material/RDF/Schema/Descriptor/midesc#"
xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#"
<midesc:Staff rdf:about="">
  < dct:isVersionOf rdf:resource="gauss1.shtml" />
</midesc:Staff>
<iwi:PersonalHomePage rdf:about="gauss1.shtml">
  < dct:isPartOf>
    <iwi:Project rdf:about="uri:x_mainid:03GRMIB4"/>
  </ dct:isPartOf>
  < dc:creator>
    < rdf:Description>
      < rdf:type>&iwi;Person</ rdf:type>
      < vCard:EMAIL rdf:parseType="Resource">
        < rdf:value>Gauss@email.www</ rdf:value>
      </ vCard:EMAIL>
      < vCard:N>
        < rdf:Description>
          < vCard:Family>Gauss</ vCard:Family>
        </ rdf:Description>
      </ vCard:N>
    </ rdf:Description>
  </ dc:creator>
</ iwi:PersonalHomePage>
</ rdf:RDF>

```

Fig. 2: XML serialization of the RDF graph

3.2 The Portal: The Math&Industry server for special services

Even the project websites have their own value, the value of their information can be increased while building further services on top of the provided data. The platform for these web based services is a portal. But the portal is more: The portal

- provides information about the funding program (funding periods, topics of the funding periods, etc.),
- hosts the project data based services,
- is the central access to the project websites,
- enables the efficient access of web crawlers to them
- and, of course, informs about the Math&Industry project itself.

The foundation of the services are the data collections of the projects and a web crawler. For this we used the Harvest⁸ software. It allows us to restrict the crawling to the relevant webpages. Furthermore it is possible to include evaluations, which are named in the crawler terminology as summarizers, for our RDF/XML pages.

Till the end of the project we set up four services at the portal: the full text search, project lists, a glossary and an expert database. and there are plans for further services, like a software-, a publication- and a product-database and research maps.

3.3 Generalisation of the concept

After reviewing our concept and results it becomes clear that there are several directions to generalise the approach of Math&Industry.

⁸ Harvest 1.5, see <http://harvest.sourceforge.net>

At first there are more projects in applied mathematics in Germany than the ones funded by the BMBF. There are at least the big players BMWi⁹, DFG¹⁰ (especially Matheon¹¹), Fraunhofer-Gesellschaft¹² and Max-Planck-Gesellschaft¹³. But there are also some smaller but nevertheless important research institutions like WIAS¹⁴. It would be a big step forward for the information supply in the sciences, industry and services, to cover all research projects in the field of applied mathematics in Germany and to give access to it by only one portal¹⁵.

Furthermore it is possible to extend the language space of Math&Industry. Three points are to consider:

1. First it should be possible to generate a project website in alternative languages to German (and of course bi- and multi-language websites should be possible).
2. Second the WebSiteMaker should present its user interface in different languages too (independent from the language of the generated websites).
3. Finally the central portal needs to support several languages - at least English.

With such an internationalisation on the software side, it would be possible, and desirable, to extend the scope of the project to the international scientific community - at least to the European Union.

But there is further potential in the Math&Industry concept, because it can be applied to other research projects beyond the field of applied mathematics. Beginning in the year 2006, we applied the Math&Industry concept to the field of "renewable energies and efficient usage of energy"¹⁶ - another funding area of the BMBF. It was the proof for the flexibility of our concept in the areas groups, subgroups, layout / design (of generated pages and tools) and subgroup classes. The change of the focus from applied mathematics to energy in ecology relations could be seamlessly carried out. The six-group-pattern of Math&Industry was replaced by a pattern with seven groups and no further subgroup types were necessary. A bigger switch was needed for the webpage generation, because the resulting pages were no longer websites on their own but part of the NGEE-portal, framed on all four sides with project independent parts.

4. CERIF and Math&Industry

Lets recall the CERIF model: First in 1991 the European Working Group on Research Databases,

9 Bundesministerium für Wirtschaft und Technologie, <http://www.bmwi.de>

10 Deutsche Forschungsgemeinschaft, <http://www.dfg.de>

11 <http://www.matheon.de>

12 <http://www.fhg.de> with its wide spectrum of institutes, especially SCAI (<http://www.scai.fraunhofer.de>) and ITWM (<http://www.itwm.fraunhofer.de>)

13 See <http://www.mpg.de/forschungsgebiete/CPT/IMK/> for the research area "Informatik / Mathematik / Komplexe Systeme"

14 Weierstrass Institute for Applied Analysis and Stochastics, <http://www.wias-berlin.de>

15 Some of them built their own information systems like GEPRIS of the DFG (<http://gepris.dfg.de/gepris/>) or present their projects at least in a systematical way like the Fraunhofer-Institute SCAI (<http://www.scai.fraunhofer.de/33.0.html>).

16 Netzwerke Grundlagenforschung erneuerbare Energien und rationelle Energieanwendung, <http://www.ngee.de>

followed by ERGO¹⁷ 1998, and now euroCRIS¹⁸ (European Professional Association of Current Research Information Systems Experts) developed CERIF, the Common European Research Information Format. It is a data model for the unified description, the data exchange and the access to research information in Europe.

The idea is simple. CERIF has four research entities on the top level: Project, Person, OrganisationUnit, and ResultPublication. These entities can be described in detail by other entities. CERIF has all in all four types of entities. Besides the top level (Core Entities) the (complex) entities like 'ResultProduct' or 'FundingProgramme' define the 2nd level entities. In addition there are 'Link Entities' to describe relations between "level entities", and as the fourth type there are 'Classification Entities' to allow the application of standard vocabularies.

The main focus of euroCRIS with CERIF¹⁹ is to provide mainly administrative data and further selected information. The aims and concepts of the projects will not be presented (within the CERIF model it is only possible to give a rough idea about the thematic and scientific content of a project by title, abstract, keywords and classification).

Both approaches provide information about research but the focus of the approaches is different. Math&Industry is project-centered, CERIF on more general information about research. In more detail:

- Both models provide “semantic relationships between research entities” (euroCRIS 2006, 3). To enable “communication and data exchange across applications” (euroCRIS 2006, 3) both models have a XML representation. Whereas CERIF build its own schema, Math&Industry uses RDF/XML.
- Both make extensive usage of foreign standards and classification systems, like Dublin Core or the ISO 639-1 two-letter language code. Especially language information is added to most of their data bits. But currency-codes are not part of Math&Industry, because until now no money related information is covered (the focus of Math&Industry lies on the research results and not on administrative data). Also country codes are not used in Math&Industry because it applies only to German projects. This will change if it will be extended to an international focus.
- In Math&Industry publications are only presented if they are a result of a project. CERIF covers also project independent publications from persons and organisations. Additional in Math&Industry an organisation can not be the author of a publication.
- While describing administrative data and more “real” things like publications, patents, persons and so on both approaches are equivalent. CERIF provides some entities which are not part of Math&Industry (e.g. research facilities and CV's) and vice versa (e.g. Software and Links). In this area a mapping of one model to the other seems possible with smaller adoptions on both sides. But while describing scientific content, CERIF can provide it only on the abstract level and with pointers to publications. Math&Industry provides a detailed schema²⁰ to cover such information within its groups midesc:ProblemInPractice, midesc:Modelling, midesc:MathematicalTreatment and midesc:Glossary. For example in chapter 3.1 we showed a small part of a RDF graph of a persons description. In the whole graph there are entities like address items vCard:Orgname²¹, vCard:Orgunit, vCard:Street,

17 <http://cordis.europa.eu/ergo/>, European Research Gateways On-Line

18 <http://www.eurocris.org>

19 as well www.forschungsportal.net and FÖKAT for example

20 see <http://www.iwi-iuk.org/material/RDF/Schema/Descriptor/midesc.html>

21 for the vCard vocabulary see <http://www.imc.org/pdi/vcard-21.txt>; the other in this paragraph used

vCard:Pcode, vCard:Locality, vCard:Country, vCard:EMAIL. All of them as “current” and “valid during project”. Always current are vCard:voice, vCard:fax, vCard:cell, vCard:TITLE (as “academic position” and “business position”). Then there are the items to describe a persons name: vCard:Family, vCard:Given, vCard:FN, vCard:Prefix, vCard:Suffix. and further there are vCard:PHOTO, vCard:ROLE, and typed links to external pages: mnst:Articles, mnst:PreprintsPublications, mnst:Preprints, iwi:Publications, mnst:CourseInformationAndMaterials, mnst:ResearchGroups, mnst:Projects, mnst:ClassSchedules, iwi:PersonalHomePage, mn:MPRESS, mn:ZblMath. Furthermore it is possible to identify the areas of interest of a person not only by keywords (dc:subject), but also to name classes from some international used classification systems. Not only the class codes are stored, but only the full label (from the root to the leaf of the schema tree). And it is possible to make statements about expert knowledge using iwip:hasKnowledgeAbout pointing to a class of our classification scheme for expertise areas. And at last there is a dc:description as place to store all other information someone would like to know others about her or him.

- CERIF uses link entities. All relations created by link entities are bi-directional. In Math&Industry every relation is mono-directional. Every 3rd-level entities (subgroups) are bound via a dct:isPartOf-relation to the project. The project does not know anything about it's parts. The same would happen with the 2nd-level entities – but they are existing only as a kind of container. No information is bound to the 2nd-level types (= groups).

Below (fig. 3) is an ER-diagram of the Math&Industry model. Compared to the CERIF it seems much less complex. But this figure shows not the whole model. The subgroups (entities of level 3) are very detailed structures. The granularity (as shown in the above example) goes far beyond the possibilities of CERIF.

Additional it should be noted that there are no relation entities in the Math&Industry model. All relations are just attributes of an entity. They are semantic links to other entities. Just to make the graphical representation somewhat comparable to the CERIF model the relations to the project entity are displayed in a link entity manner.

As a result we can say that the Math&Industry concept is more comprehensive than the CERIF approach. More technically, the entities of CERIF can be seen as a subset of the Math&Industry data model.

To come to a conclusion the question of a mapping or a linking must be answered. Theoretically, a mapping seems possible in both directions. But a mapping from CERIF to Math&Industry can, of course, only applied for research projects. But before doing this, both models must be extended to fill the entity gaps. And while doing this on the CERIF side much more things are to do compared to the needed extensions for Math&Industry.

Though the scope of CERIF is more general it would be a nice solution to set up a RDF-Schema for CERIF. Then it would be possible for Math&Industry to use the CERIF-vocabulary and to link all other items to it with isSubClassOf- or seeAlso-relations. CERIF could be the “European Core” of research metadata while open itself for extend it locally and still be interoperable to exchange research information.

namespaces / schemas are 'mnst' (<http://www.iwi-iuk.org/material/RDF/1.1/descriptor/mnst.rdf>), 'mn' (<http://www.iwi-iuk.org/material/RDF/1.1/Schema/Class/mn#>), 'dc' (<http://purl.org/dc/elements/1.1/>), 'iwip' (<http://www.iwi-iuk.org/material/RDF/Schema/Property/iwip#>)

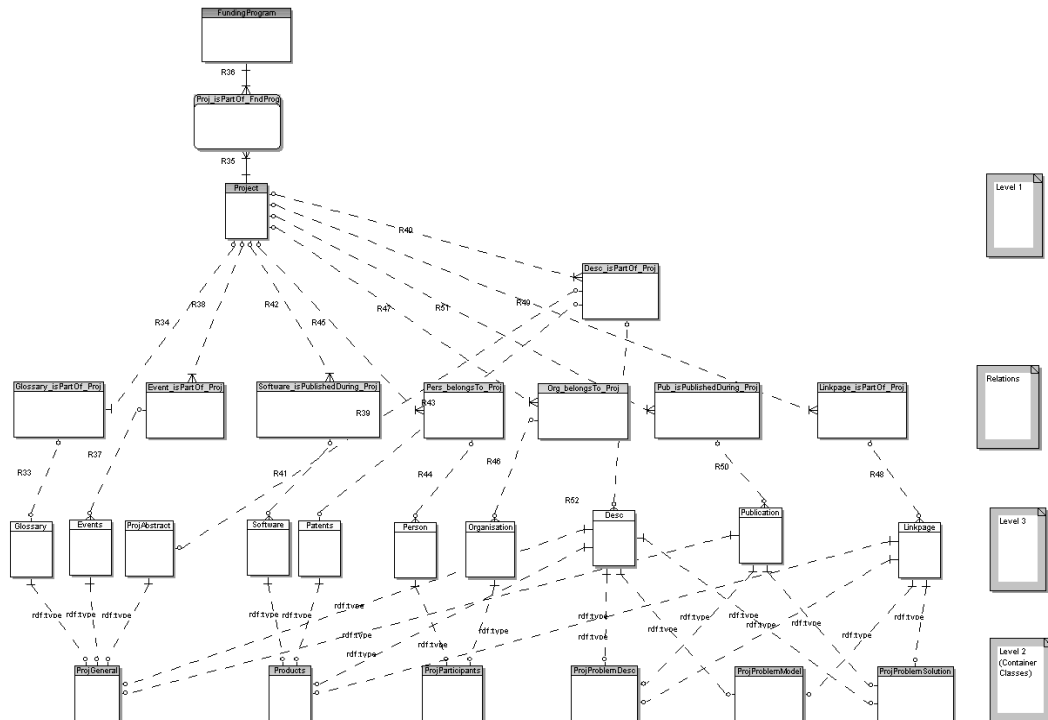


Fig. 3: The (reduced) Math&Industry model as an ER-diagram

5. Outlook

There would be some things on the Math&Industry-agenda to do. We already mentioned the generalisation perspectives.

Additionally we developed concepts of further services: databases for software, publications and products, automatically built classifications of applied mathematics (one of used mathematical methods and one of the application areas), interactive research maps, an extension of the search facility by the method of faceted browsing (see for example Allen 1995).

Even the Semantic Web did not evolve as fast as Tim Berners-Lee expected in 2001, it is a useful invention with a huge potential for the future. Grounding our concept on RDF/XML we showed this. It fits well to the needs of an open scientific community where the problematic layer of trust can be ignored²². It can be ignored because within our structure there are no interests to falsify statements or to mistrust the sources. So, more in the background, but an important extent for the Semantic Web research, would be to switch from the parsing of the RDF/XML-files by the tools which created them to real RDF triple based evaluation and using the query language SPARQL²³ to access the semantics without the need to know how the XML-structure of the RDF-coding

²² More about semantics in the Internet and Math&Industry, see the (Roggenbuck 2006)

²³ see <http://www.w3.org/TR/rdf-sparql-query/>

looks like. As a further step to make the XHTML-presentation more flexible it seems to be promising to generate it using server sided XSLT. Doing so will result in an online generation of the presentation for men and free the WebSiteMaker of managing the presentation.

These perspectives of further development may be part of another project which takes up the loose ends of these development- and research-threads. The Math&Industry project ended in July 2007. Some maintenance is assured till the end of 2008.

References

Allen, Robert B. (1996): *Retrieval from facet spaces*. In: Electronic Publishing, 247-257, 8, 1995. (presented as Electronic Publishing 1996, Palo Alto, CA). <http://www.ischool.drexel.edu/faculty/ballen/PAPERS/FACETS/facets.html>

Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001): *The Semantic Web*. In: Scientific American, 29-37, May 17th

euroCRIS (2007): *CERIF2006-1.1 Full Data Model (FDM) – Model Introduction and Specification*, October 17th (http://www.dfki.de/~brigitte/CERIF/CERIF2006_1.1FDM/CERIF2006_FDM_1.1.pdf)

Mills, Elinor (2005): *Google to Yahoo: Ours is bigger*, CNET News.com, September 26th, http://news.com.com/Google+touts+size+of+its+search+index/2100-1038_3-5883345.html

Roggenbuck, Robert (2006): *"Web Site Erstellung - was leisten Content Management Systeme?"*, talk on the annual conference of the Deutsche Mathematiker Vereinigung 2006 in Bonn, given in the minisymposium "Information, Kommunikation und Bibliotheken für die Mathematik", September 19th, <http://www.mathematik-21.de/publications.shtml#pub060919>

Contact Information

Robert Roggenbuck
Institut für wissenschaftliche Information Osnabrück e.V. (IWI)
Albrechtstr. 28a
49069 Osnabrück
Germany

e-mail: robert.roggenbuck@uni-osnabrueck.de