GES 2007
German e-Science

Available online at http://www.ges2007.de

An Architecture Blueprint for Knowlege-based e-Science[*]

Claudia Niederée[1], Thomas Risse[1], Marco Paukert[2] and Adelheit Stein[2]

[1] Forschungszentrum L3S, Appelstr. 9a, 30167 Hannover, Germany
email: {niederee, risse}@l3s.de
[2] Fraunhofer IPSI, Integrated Publication and Information Systems Institute,
Dolivostrasse 15, 64293 Darmstadt, Germany
email: {paukert, stein}@ipsi.fraunhofer.de

1 Introduction

The *scientific innovation process* embraces the steps from problem definition through the development and evaluation of innovative solutions to their successful exploitation. The challenges imposed by this process can be answered by the creation of a powerful and flexible next-generation e-Science infrastructure, which exploits leading edge information and knowledge technologies and enables a comprehensive and intelligent means of supporting this process.

For such an infrastructure, there are various areas with a potential for improved innovation process support like:

- a more effective re-use of scientific results, information and various other kinds of "innovation" resources
- facilitating the collaboration in dynamically created multidisciplinary teams
- an accelerated and more effective fostering of technology transfer
- intelligent support for routine tasks enabling the researcher to focus on the creative parts of innovation
- a flexible management of the innovation process, which takes into account the dynamics and the creative character of this process.

However, for creating a successful and widely accepted e-Science infrastructure it is crucial to get acquainted with the needs and working practices of the actors involved in scientific innovation processes as well as to take into account the human and organizational context of innovation.

The construction of a *Knowledge-based e-Science Infrastructure* (KeSI) does not need to start from scratch. Past and current developments in the area of digital libraries, Grid technologies, and knowledge technologies, for example, provide important building blocks for the construction of a KeSI.

The next section presents the results of an in-depth study of the researchers' requirements. Section 3 describes our vision of a Knowledge-based e-Science infrastructure. A blueprint architecture is presented in Section 4. Afterwards Section 5 gives an overview about the *Fraunhofer e-Science Cockpit* as a first implementation of our vision. A discussion of the state of the art and contributing

[*] This work has partly funded by the German Minstry of Education and Research (BMBF)

technologies is conducted in Section 6. Finally, the paper concludes and gives an outlook on open issues.

2 What do Researchers Request?

The Fraunhofer e-Science Cockpit is planned to be established as a supporting system for researchers within Fraunhofer-Gesellschaft. As one of the leading organizations for applied research in Europe its 56 research institutes cover a variety of research areas, focussing on engineering fields rather than on basic research or mainly natural science such as the Max-Planck-Gesellschaft.

In order to review our e-Science vision from a practical viewpoint we had to investigate the real-life requirements of potential users/researchers. To this end, two comprehensive surveys with standardized questionnaires were carried out.

The first study was conducted in 2004, addressing *all non-universitarian, German research organizations* (for detailed survey results see [15]). The institutions were to describe their use of internal and external information and data resources, their current computer environment and IT-infrastructure, and their assessment of e-Science. Whereas the majority (70%) was interested in actively participating in future e-Science projects, only part of them was hereby referring to large-scale computing resources and activities. These occur predominantly in natural science and to a smaller degree in technical engineering fields. Institutions with no such 'traditional' e-Science applications still expressed a strong interest in a knowledge, information and collaboration based platform for distributed research communities (e.g. a "collaboratory", cf. [8, 13]).

The second survey was carried out in 2006, now addressing *all individual researchers of the Fraunhofer-Gesellschaft*. For this special purpose a more fine-grained questionnaire was developed. Since the design of the questionnaire, the survey procedure and data analysis were performed according to professional methods of empirical research, we do have now at hand a rich basis of sound empirical data from the researchers themselves. 869 respondents returned the questionnaire. With respect to its composition the random sample is representative to our target population (21% of about 4.200 Fraunhofer researchers).

The questionnaire contained 30 complex questions split into 150 single items for each of which the respondents were to indicate their grade of agreement or disagreement on multiple point scales. These were complemented by some open-ended questions for additional items and commentaries.

The respondents give a detailed account of their personal assessment of all facets of the *scientific innovation process*, such as (a) usage of information resources, (b) work with scientific data, (c) use of scientific software, methods, computational resources, (d) difficulties with new technologies and facilitating measures, (e) use of publication tools and support (for the detailed findings see [14]). Combining all explicit and implicit (statistically derived) evidence from our data analyses we can roughly categorize the conveyed user requirements into five main fields, ordered according to their relevance.

Search and retrieval of scientific/technical information ranks on top of this list, as most of the *explicitly* formulated requests for more/better IT support can be subsumed here. The researchers' "wish lists" range from easier, more comfortable, free access to existing information resources (digital libraries, Intranet, etc.) over better, richer, value-added information (including evaluated data and information, recommendations) up to better search functionalities (e.g. semantic and context-based methods), and intelligent integration into the work environment. The available information resources are used by the majority, but quite infrequently (except for Internet search engines, which are used by 85%). About 20-25% respondents judge critically about current day information provision systems, and mention a number of concrete problems, e.g., 60% respondents are not sure whether and how often they miss important information.

Collaboration with colleagues and virtual communities is a very important topic all over the questionnaire, since most research tasks can be tackled individually or collaboratively. However, collaboration as such is not considered as problematic or deficiant, hence requests for more IT support sound less acute or insistent. Some respondents utter concrete requests for help in finding relevant contact partners/experts in a given (new) field or for support of collaborative document management in virtual teams. The willingness of many respondents to share their own experimental data with external colleagues and communities is surprisingly high. Here, we find a big potential for more IT support, especially when collaborative *work* with data and other resources is in the question.

Working with scientific/experimental data appears to be less extensive in the work of Fraunhofer researchers than in other – basic research oriented – institutions, e.g. in the natural sciences (although exact comparative data do not exist). About 20% of the Fraunhofer researchers spend more than 40% of their time with data collection and analysis, some of them using huge data sets. Half of the respondents spend up to 20% and one fourth up to 40% of their time. Thus, working with data is no hot topic, but two third of the respondents still would prefer more IT support.

Publishing tools and **IT infrastructure assistance** often play an important role in everyday work, but we have no evidence that additional IT support is required. The variety of employed text and graphics editing tools is amazing, but no widely significant problems are reported here aside from a few, very concrete suggestions for improvement, mostly on the organizational level.

Most of the findings support our knowledge and collaboration-based vision of e-Science impressively. The detailed survey results will be further exploited carefully in order to allow for a demand-driven, user-centered implementation of the Fraunhofer e-Science Cockpit.

3 Knowledge-based e-Science

In this section we develop a vision for a knowledge based e-Science infrastructure by identifying possible development directions for a KeSI based on the study results and a state of the art analysis.

Currently, strongest support for the working processes of the researcher can

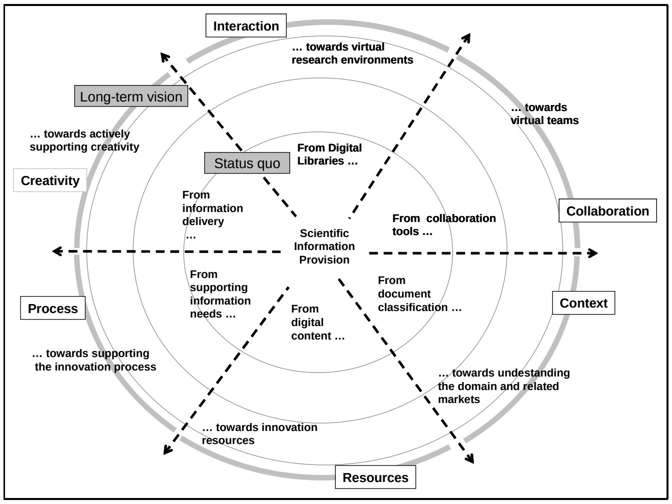

Interaction
... towards virtual research environments
Long-term vision
... towards virtual teams
... towards actively supporting creativity
Status quo
From Digital Libraries ...
Creativity
From information delivery ...
Collaboration
From collaboration tools ...
Scientific Information Provision
From supporting information needs ...
From document classification ...
Context
Process
From digital content ...
... towards supporting the innovation process
... towards undestanding the domain and related markets
... towards innovation resources
Resources

Figure 1: R&D Directions for next generation e-Science Support

be found in the area of easing the access and use of relevant information sources (*information provision*). This includes advanced generic search services (based on information retrieval as well as fielded search in metadata records), support for accessing application specific information and data collections, as well as services for the structuring, enrichment, management, and preservation of scientific documents, typically for a target scientific community.

Taking typical scientific information access support by a digital library as a starting point, six directions of further development for more comprehensive support of the scientific innovation process - as targeted by knowledge-based e-Science support - have been identified (see also figure 1):

Resources: Here a systematic transition from the systematic management and dissemination of digital content (as it is done in DLs) to the management of various types of innovation resources is desirable. [2] For this purpose, formats and annotation processes for the adequate description of the resources by metadata as well as services for their effective and access are required, which takes into account the process context.

Context: In this R&D direction the goal is an improved understanding of the domain. The domain and the domain knowledge provide the context for the work of the scientist. In this case, a transition from the basic structuring of the domain, as it is, for example, achieved by thematic document classification, to services that support a deepened understanding of the domain, its structure, trends and the associated markets is targeted.

[2] Scientific documents and other digital content can be considered as a special type of innovation resources. Other types of innovation resources are services, electronic and physical tools, expertise, scientific methods, in summary all types of resources that are used in the scientific innovation process.

Collaboration: Various collaboration support tools and services already exist that can be exploited in scientific collaboration. Therefore the aim of this direction is to enable easy and demand-oriented access and activation of such tools to provide adequate collaboration support depending on the respective situation in the innovation process. Further requirements are imposed for efficiently supporting multidisciplinary teams (e.g. bridging terminology differences) and for enabling the dynamic formation of virtual research teams.

Interaction: In digital libraries, the interaction is dominated by the typical interaction with search tools and further functionality in support of information seeking behavior [9]. Here a transition to richer interactions with the user in the spirit of virtual research environments is targeted. This includes intelligent services that can perform routine tasks for the researcher, services that enable the researcher to act as an information provider (not only as a consumer), as well as the integrated use of new forms of user interfaces (haptic interfaces, voice control, smart boards, support for the mobile researcher).

Creativity: Effective IT support for creativity, which is in the core of innovation, is definitely the most challenging part in developing a knowledge-based e-Science infrastructure. A starting point is the evaluation and adaptation of creativity tools like mind mapping and brain storming support. Indirect forms of creativity support include tools and services for more effectively learning from the experiences of others and for checking the innovativeness as well as the feasibility of new ideas.

Process: Here a transition from only supporting information access (and management) towards a more comprehensive and integrated support of the activities in the innovation process is desirable. This can be achieved by an extensible e-Science environment that integrates generic as well as application specific tools and services supporting innovation activities. In addition the innovation process itself can be modeled and monitored in the sense of a workflow. However, the flexible and highly dynamic character of this process has to be taken into account.

An effective knowledge-based e-Science infrastructure will have to address challenges in all these six directions. However, it makes sense to identify a set of core functionalities and then to step-wise extend the infrastructure in close collaboration with the research communities, ensuring demand-driven evolution.

4 Architecture Blueprint

In order to implement the e-Science vision described in the previous sections, we propose a generic reference architecture blueprint for a knowledge based e-Science infrastructure depicted in Figure 2. Due to the complexity of the required technologies we decided to group the services and components into four layers.

Resource Network Infrastructure: The aim of the Resource Network Infrastructure layer is to provide access in a standardized way to different

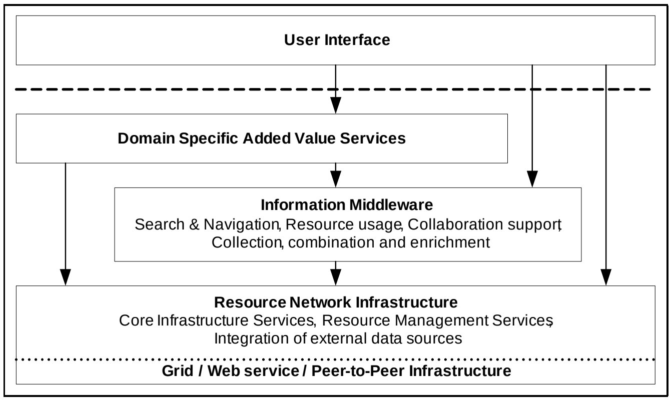

User Interface
Domain Specific Added Value Services
Information Middleware
Search & Navigation, Resource usage, Collaboration support,
Collection, combination and enrichment
Resource Network Infrastructure
Core Infrastructure Services, Resource Management Services,
Integration of external data sources
Grid / Web service / Peer-to-Peer Infrastructure

Figure 2: Architecture Blueprint for an e-Science Infrastructure

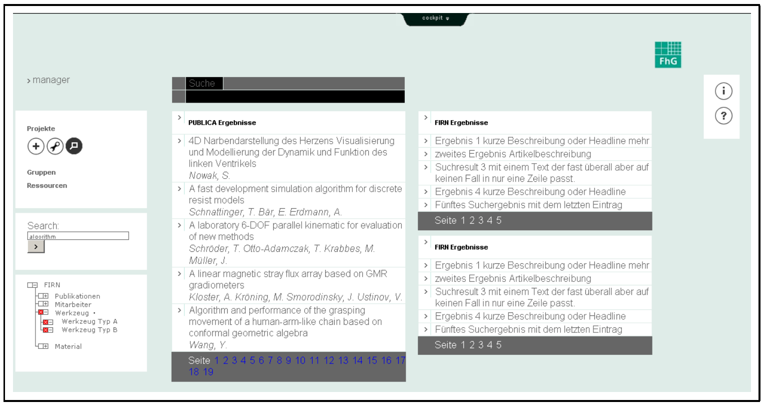
resources like digital libraries, computational services or storage capacities. The obvious architectural approach are services as they are implemented by web services or grid services. The selection of a technology depends on the requirements of the respective scientific domain. However, to achieve the goal of availability and sustainability, a service oriented architecture should be complemented with self-organizing infrastructure services. The reason is that centralized environments have a higher possibility of failure. Self-organizing decentralized environments can act very flexible on changes in the infrastructure (e.g. failure of a node). An example for a self-organizing Web service infrastructure is the BRICKS framework[3].

Information Middleware: The Information Middleware layer aims to provide generic services for the interaction between innovators and between innovators and innovation resources as well as for handling the different types of innovation resources in a unified way. This layer is independent from concrete applications. The information middleware consists of four functional groups.

Search and navigation has to support the integrated access to various heterogeneous resource repositories like digital libraries or service repositories. To make this functionality as effective as possible for the researcher, the current working or innovation process context have to be taken into account.

The second service group supports the *re-use of innovation resources*. This includes services for the publication of resources, for monitoring their usage, and for accessing resources. Furthermore, this service group also includes services for the rating of resources and notification services, which notify the researcher about certain resource-related events.

In order to support *collaboration* between researchers the collaboration services provide base functionalities required by a number of applications. An important collaboration service is the annotation service that allow researchers e.g. to annotate a service with usage experiences or to discuss



PUBLICA Ergebnisse

- 4D Narbendarstellung des Herzens Visualisierung und Modellierung der Dynamik und Funktion des linken Ventrikels *Nowak, S.*
- A fast development simulation algorithm for discrete resist models *Schnattinger, T. Bär, E. Erdmann, A.*
- A laboratory 6-DOF parallel kinematic for evaluation of new methods *Schröder, T. Otto-Adamczak, T. Krabbes, M. Müller, J.*
- A linear magnetic stray flux array based on GMR gradiometers *Kloster, A. Kröning, M. Smorodinsky, J. Ustinov, V.*
- Algorithm and performance of the grasping movement of a human-arm-like chain based on conformal geometric algebra *Wang, Y.*

Seite 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

FIRN Ergebnisse

- Ergebnis 1 kurze Beschreibung oder Headline mehr
- zweites Ergebnis Artikelbeschreibung
- Suchresult 3 mit einem Text der fast überall aber auf keinen Fall in nur eine Zeile passt.
- Ergebnis 4 kurze Beschreibung oder Headline
- Fünftes Suchergebnis mit dem letzten Eintrag

Seite 1 2 3 4 5

FIRN Ergebnisse

- Ergebnis 1 kurze Beschreibung oder Headline mehr
- zweites Ergebnis Artikelbeschreibung
- Suchresult 3 mit einem Text der fast überall aber auf keinen Fall in nur eine Zeile passt.
- Ergebnis 4 kurze Beschreibung oder Headline
- Fünftes Suchergebnis mit dem letzten Eintrag

Seite 1 2 3 4 5

Figure 3: Navigation screen with search results

the content of a publication.

Finally, the Information Middleware provides functionalities that allow the *collection of information and resources* for specific innovation tasks, the *enrichment of such resources* and their combination with other information and innovation resources into structured collections e.g. in preparation of a publication.

Domain Specific Added Value Services: The Domain Specific Added Value Services compose and enhance the information middleware services in order to provide domain specific support. Furthermore, this layer also provides application specific services and service extensions. Hence in the design of this layer it is necessary to involve the research community to a high degree.

User Interfaces: The User Interfaces are typically domain specific, but generic components are possible. The design of an adequate user interface is challenging, as on the one hand it has to be easily understandable but on the other hand must offer the rich functionality of the KeSI to the user.

5 Fraunhofer e-Science Cockpit

A prototype implementation of the previously introduced blueprint of a knowledge based e-Science architecture has been implemented within the Fraunhofer e-Science Cockpit project. The aim of the project is to support the researcher in their scientific innovation process in applied research. The idea is based on the metaphor of a “cockpit”. Hence, the researcher should be able to navigate through the space of innovation resources. The cockpit provides an integrated and context oriented access to information sources of scientific relevance (i.e. scientific digital libraries) and market relevance (i.e. market studies, collections of patents).

In addition, the cockpit allows a flexible, demand-oriented usage and systematic re-use of available innovation resources. This helps to reduce delays and therefore costs by making the innovation process more efficient. Another important aspect is the seamless and traceable integration of scientific data, e.g. into the publication process. This enables an easier verification and comparison of scientific results, as data is enriched with semantic information on how it was created, aggregated, and interpreted.

Collaboration is another important part of the daily scientific work. Therefore, the ad-hoc creation of interdisciplinary project communities – enabling the sharing of knowledge and resources – will be supported by the system.

Technologically, the whole system is based on a decentralized and service oriented framework developed within the BRICKS project[3]. The BRICKS framework integrates a number of base services, (e.g. management of content and metadata, federated search, handling of security), while being very flexible in the customization for different kinds of applications and user domains.

The information middleware implements the services as described in the architecture blueprint. The innovation resources are modeled as RDF/OWL ontologies. The challenge was to develop a flexible resource model, which allows the flexible addition of new innovation resources while respecting all relevant aspects of resource descriptions. The selected approach describes the different facets of a resource through separate profiles. This allows the domain specific extensions of resources and the usage of domain specific names for the attributes of a resource.

In the first version of the cockpit the user interface is based on state-of-art Web 2.0 technologies. Figure 3 shows the results of a query, which can be further refined by deselecting resources within resource tree on the left side of the screen. For later stages of the project it is planned to integrate the functionalities directly into the daily used applications of a researcher.

6 Related Work and Contributing Technologies

Complementary to the german D-Grid[1] initiative, the German Minstry of Education and Research (BMBF) is funding a portfolio of projects, which are aiming towards knowledge-based e-Science as it is discussed in this paper. Two of them, the project eSciDoc and the project FRESCO (cf. Section 5) are dedicated to support the scientific innovation process in the two large German research organizations, the Max Planck Society (MPG) and the Fraunhofer Gesellschaft, respectively. The aim of the eSciDoc project [2] is to support the publication process and improving the information flow by developing a platform for communication and publication for research organizations. Another example is the project Ontoverse, which builds an infrastructure for the collaborative and multidisciplinary construction of ontologies with a special focus on life sciences.

A prominent example from the UK e-Science programme is myGrid [11]: it offers a semantic grid system for the bio-informatic community. In this project, semantic web technologies are used to improve the service and resource discovery by enhancing their descriptions with ontologies and reasoning on top of them.

Contributing Technologies

The construction of a knowledge-based e-Science infrastructure can build on a number of past and current developments in the area of digital libraries (DL), Grid technologies, and knowledge technologies.

Current development trends in DL architectures are aiming for a transition from the DL as an integrated, centrally controlled system to a dynamic configurable federation of DL services and information collections [7]. This transition is inspired by new technologies like Web services, the grid as well as peer-to-peer networking. Activities which are working in these directions are for example DSpace [10], which captures, stores, indexes, preserves and redistributes an organization's research material in digital formats. In LibraRING [16] the goal is to setup a completely decentralized infrastructure of distributed digital libraries. A similar approach is taken in the BRICKS project[3], but it is providing a richer set of functionalities. Another example in this direction is DILIGENT[5], which aims to provide a DL as a dynamic grid resource.

Semantic Web technologies have a clear application in an e-Science infrastructure, especially by supporting the more effective re-use of innovation resources. Semantic Web Research areas that are relevant for the e-Science context include methods and tools for ontology engineering (e.g. [4]) as well as semantic search (e.g. [6]), which uses semantic information to improve the search process and the query results. For putting the Semantic Web in operation, automated pipelines for the creation of semantic information are required (e.g. [12]), since manual rich annotation of large amount of content is too expensive. The establishment of annotation pipelines is related to work on supporting provenance workflows [17], where metadata on scientific data and the steps leading to its creation are automatically captured.

7 Conclusion and Outlook

In this paper we first analyzed the requirements of researchers based on a study conducted within the Fraunhofer Gesellschaft. These results helped to refine and verify our overall vision for knowledge based e-Science. Driven by this vision we developed an architecture blueprint for the construction of a knowledge-based e-Science infrastructure. As a concrete example, the blueprint provided the basis for the creation of an e-Science infrastructure for the Fraunhofer Gesellschaft in the FRESCO project. Overall, the chosen approach seems to be promising for future developments in better supporting researchers' daily work. Even though varous partial solutions already exist from different projects, a there are still major research, development and integration challenges for a real integrated portfolio of tools that effectively support researchers.

References

1. BMBF. *D-GRID Initiative*, 2006. http://www.d-grid.de/.
2. BMBF. *eSciDoc*, 2006. http://www.escidoc-project.de/.

3. BRICKS Project. *BRICKS - Building Resources for Integrated Cultural Knowledge Services (IST 507457)*, 2004. http://www.brickscommunity.org/.
4. Oscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez, editors. *Ontological Engineering.* Springer-Verlag London Limited, 2004.
5. DILIGENT Project. *DILIGENT - A testbed DIgital Library Infrastructure on Grid ENabled Technology*, 2004. http://www.diligentproject.org/.
6. J. Hartmann, N. Stojanovic, R. Studer, and L. Schmidt-Thieme. Ontology-based query refinement for semantic portals. In M. Hemmje, C. Niederée, and T. Risse, editors, *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, volume 3379 of *Lecture Notes in Computer Science*, pages 41–50. Springer, 2005.
7. P. Knezevic, B. Mehta, C. Niederée, T. Risse, U. Thiel, and I. Frommholz. Supporting information access in next generation digital library architectures. In C. Türker, Ml Agosti, and H.-J. Schek, editors, *DELOS Workshop: Digital Library Architectures*, volume 3664 of *Lecture Notes in Computer Science*, pages 207–222. Springer, 2004.
8. R.T. Kouzes, J.D. Myers, and W.A. Wulf. Collaboratories: Doing science on the Internet. *IEEE Computer*, 29(8):40–46, 1996.
9. Lokman I. Meho and Helen R. Tibbo. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society of Information Science and Technology*, 54(6):570–587, April 2003.
10. MIT Libraries & Hewlett-Packard. *DSpace.* http://www.dspace.org/.
11. Open Source Project. *myGrid: Middleware for in silico experiments in biology*, 2006. http://www.mygrid.org.uk/.
12. Rodolfo Stecher, Claudia Niederee, Paolo Bouquet, Thierry Jacquin, Salah Ait-Mokhtar, Simonetta Montemagni, Roberto Brunelli, and George Demetriou. Enabling a knowledge supply chain: From content resources to ontologies. In *Workshop "Mastering the Gap" on European Semantic Web Conference (ESCW'06)*, June 2006.
13. A. Stein, J. Keiper, L. Bezerra, H. Brocks, and U. Thiel. Collaborative research and documentation of European film history: The COLLATE Collaboratory. *International Journal of Digital Information Management (JDIM), special issue on Web-based Collaboratories.*, 2(1):30–39, 2004.
14. A. Stein and M. Paukert. Anforderungen der Fraunhofer-Mitarbeiter an eine verbesserte IT-Unterstützung der Forschung: Ergebnisse einer Fragebogenerhebung (88 p.). Technical report, Fraunhofer IPSI, Darmstadt, 2006. http://www.ipsi.fraunhofer.de/ stein/publications.html.
15. Adelheit Stein. Anforderungen zukünftiger Nutzer einer eScience-Infrastruktur. Ergebnisse einer Fragebogenerhebung (47 p.). Technical report, Fraunhofer IPSI, Darmstadt, 2005. http://www.ipsi.fraunhofer.de/ stein/publications.html.
16. C. Tryfonopoulos, S. Idreos, and M. Koubarakis. LibraRing: An Architecture for Distributed Digital Libraries Based on DHTs. In Andreas Rauber, Stavros Christodoulakis, and A. Min Tjoa, editors, *ECDL*, volume 3652 of *Lecture Notes in Computer Science*, pages 25–36. Springer, 2005.
17. FJun Zhao, Chris Wroe, Carole Goble, Robert Stevens, Dennis Quan, and Mark Greenwood. Using semantic web technologies for representing e-science provenance. In *The Semantic Web - ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004*, volume 3298 of *Lecture Notes in Computer Science*, page 92. Springer, January 2004.