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

eSciDoc – a Scholarly Information and Communication Platform for the Max Planck Society

Malte Dreyer[1], Natasa Bulatovic[1], Ulla Tschida[1], Matthias Razum[2]
[1] Max Planck Digital Library, Amalienstr. 33,
80799 München, Germany
{malte.dreyer, bulatovic, tschida}@mpdl.mpg.de

[2] FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1,
76344 Eggenstein-Leopoldshafen, Germany
matthias.razum@fiz-karlsruhe.de

Abstract

eSciDoc is as a joint project of the Max Planck Society and FIZ Karlsruhe, funded by the Federal Ministry of Education and Research (BMBF), with the aim to realize a next-generation platform for communication and publication in research organizations [1]. The result of the entire eSciDoc project is intended to ensure open and persistent access to the research results and materials of the Max Planck Society and to integrate these materials in an emerging e-Science network, to increase the accountability of research and to improve the visibility of the Max Planck Society. At the same time, the project aims to provide effective and comprehensive access to information for Max Planck researchers and their work groups. Additionally, eSciDoc will support scientific collaboration and interdisciplinary research in future e-Science scenarios and optimize the exploitation of information available through an interconnected global scientific knowledge space.

1 Introduction

E-Science (or “enhanced science”) is an extended way of net-based scientific working within the Scientific Information Domain. New information and knowledge technologies can help to further improve existing organizational structures, information management, and available tools, thus simplifying, intensifying, and accelerating research processes. John Taylor (former Director General of Research Council, UK) has summarized this fact as follows: “e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it”. E-Science means cooperation, collaboration, and communication throughout the whole process of knowledge generation.

E-Science demands innovative methods, services, and technology infrastructures that effectively support researchers in their daily working processes. Often, the term e-Science is used to describe computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing. This led to the deployment of large data centers and physical networks, which became basic elements of the evolving e-Infrastructures. Vast amounts of experimental and primary data are available today in scientific repositories, and these numbers are growing dramatically [2][3]. Current research is more and more data driven, even in the Humanities, which requires direct access to primary data.

However, e-Science should be understood in a broader sense. Direct access to primary data, publications, and other scholarly material does not necessarily require the deployment of large-scale computational and storage resources. More and more institutions are building up digital libraries and repositories to allow for easy, fast, and effective access to scholarly material for their scientists and students, thus improving and accelerating scholarly work. Consequently, the amount of data directly accessible to scientists is growing, which makes it increasingly difficult to keep track of newly published material and to filter out the relevant information. Additionally, there is a gap between the large scientific repositories and the existing digital libraries. The challenge is to bridge that gap and integrate all existing sources into one scientific knowledge space. This problem is even aggravated by the ongoing intensification of cross- and multidisciplinary research. Sophisticated information and knowledge management is therefore the key issue to let the vision of e-Science become reality. Knowledge management deals with the ways to influence and administer the knowledge base within an organization or institution. The knowledge base encompasses all data, information, knowledge, and competency that an organization utilizes to solve problems and achieve goals. Important steps towards a functional scientific knowledge space include the standardization of data formats, the development of well-accepted ontologies, and the management of heterogeneity in cases where standardization is not feasible.

Current information systems often concentrate on structured data. However, much of the existing scholarly and scientific material is semi- or unstructured data. Today’s systems are often restricted to types of materials, which have beforehand been curated by traditional libraries. However, scholarship produces additional types of information units, which never hit the shelves of libraries: primary data, results from simulations, informal results, findings, annotations, and so forth. These form information objects of their own right and should be adequately treated in within the scientific knowledge space [4]. They need to be citable and allow for stable references, which require long-term preservation, persistent identification, and meaningful metadata. New object types ask for new ways to search, aggregate, and visualize them. The ability to reference objects within the scientific knowledge space not only mimics the existing well-established practice of citing existing material, but also evolves it into a much richer tool that allows the creation of compound objects from various sources. Relations within the scientific knowledge space can be enhanced by adding semantics to relations, thus creating high-level or domain-specific ontologies. In such environments, provenance information and reliability of access becomes pivotal.

2 The eSciDoc Project

The eSciDoc Project, funded by the German Federal Ministry of Education and Research, aims at realizing a platform for communication and publication in scientific research organizations. eSciDoc is a joint project of the Max Planck Society (MPS) and FIZ Karlsruhe. It intends to:

- Ensure permanent access to the research results and research materials of the Max Planck Society and seamless integration within eSciDoc as well as integration into an emerging, global scientific knowledge space;
- Provide effective opportunities for access to information for scientists of the Max Planck Society and work groups;
- Support scientific collaboration in future e-Science scenarios.

The project is not only aiming at specific scientific disciplines, but has a much broader scope. The goal is to provide a generic infrastructure and solution environment for all scientific sections of the Max Planck Society (natural sciences, life sciences, social sciences and humanities). At the same time, the project takes the different needs and requirements from the sections into account and learns from existing specific solutions and projects already existing within the e-Science context at the MPS, e.g., the ECHO system [5], or the GAVO project [6].

eSciDoc is not driven by technology and computer science only. The team brings together functional requirement engineers, describing the scenarios and required use cases, as well as technological experts and developers. Both groups within the team are tightly coupled within a well-established communication and tuning process to deeply discuss the different views and to achieve commonly accepted approaches fulfilling the needs of the stakeholders within the scientific community.

The infrastructure needed for an e-Science environment has to focus on basic technologies of computation, communication, software programs, and the adequate data preservation systems needed to manage large-scale data sets consisting of multiple media, particularly those of the cultural heritage domain. To better address the challenge of realizing a generic infrastructure, which at the same time is able to support discipline-specific applications on top, the project chose an architecture that allows the modelling of modular services. For the development of solutions, i.e. community-specific viewing and working environments, these services are combined to meet the end-user-specific needs.

The first solutions currently implemented cover the fields of publication and scholarly data management. They address the basic demands for reliable depositing, retrieval, and management of different types of data within the scientific lifecycle. For working with these data and collections, flexible versioning and workflow models have to be implemented before further services for more sophisticated data processing can be supported. As these data processes are extremely domain-specific, they should be supported in a generic way, offering standardized interfaces to integrate them as externally created and maintained services.

To enable the eSciDoc infrastructure to interoperate best possible with other existing systems and data sources, adequate discovery and sustainability of the deposited or referenced data as well as different services to handle multiple metadata profiles are essential. Besides basic and widespread metadata formats, like Dublin Core or MODS, more community-specific metadata profiles for certain solutions have to be supported by the infrastructure. The developed services need to provide high flexibility in this respect, since these formats cannot be predicted in advance for all potential eSciDoc solutions,.

This agnostic approach to data and metadata has lead to the development and implementation of abstract content models, to enable us to deal with different and yet unknown types of data and structures. In this context, the project will cooperate on the the further development of Fedora, the repository system in use for basic object management.

The eSciDoc system supports relations between deposited objects, to express basic semantic associations without making any strong assumption about possible content and data structures. Further, more sophisticated semantic relations will be added to the initial ontology on demand.

3 Functional View on eSciDoc

Advances in science, scientific activities and processes will arise from unhindered international collaboration among scholars. To exempt researchers from barriers of physical location of knowledge (data) and know-how (individuals), a solid infrastructure to provide data storage, interoperability and seamless integration into community-specific working environments is necessary. The technical and technological means provided by a scientific infrastructure are fundamentals in enabling e-Science, however, the essential requirement to enable innovation and research is the adequate working environment for scholarly communities. Even within a discipline, the requirements for a specific solution might differ, be it in presentation aspects or functionalities. Therefore, the functional requirements for the eSciDoc project are threefold: provide community-specific solutions which serve the needs of a specific research group or research question, provide standardized interfaces for other communities or systems to re-use data and functionalities and provide adequate user interfaces and viewing environments where necessary.

3.1 Functional requirements

Future research and innovation builds upon the complete research output gathered during a research life cycle. This includes traditional publications such as articles or conference-papers, which are stored in interoperable archives. One of the solutions build within the eSciDoc project, the publication management, will address the complex requirements for building an institutional archive, with configurable publishing workflows and presentation views and multiple features for data management, data presentation and data integration. Still, an institutional archive will manage only one part of the research output, which represents a mere snapshot of documented findings in a steadily ongoing research lifecycle. In addition, it is limited to traditional publication models.

eSciDoc, however, is in addition aiming for solutions to manage primary data, which are gathered during the research process and mostly go beyond traditional publication models. Primary data, such as digitised images, simulation data, primary data of field studies, relations, statistical models etc. are produced in multiple ways, be it via human interaction in collaborative authoring environments, be it by re-using existing data sets in different environments or be it via data-gathering devices in data-intensive research. In difference to "solid", cohesive data, e.g. a sequence of a particular gene or a published article, research data is increasingly becoming "flowing", i.e., data entities are continuously re-used and re-combined and further modified, depending on research question and environment. The second solution currently developed in eSciDoc, the Scholarly Workbench, deals with digital objects from the arts and humanities in a collaborative authoring environment. That makes the scientific discourse a really dynamic process. At the same time, systems and functionalities have to be modelled in a way to make it easier for scientists to separate data from argumentation.

Above all, the eSciDoc project is driven by the aim of the MPS to provide free and open access to its research output. Still, the legal implications for storing and publishing research output are manifold. While the storage and dissemination of traditional publications has to follow complex policies and licenses due to the services and agreements offered by commercial publishers and learned societies, the handling of primary data has to focus in addition on potential implications due to data privacy and patent restrictions. These issues are not to be underestimated when designing functionalities and system, to be able to offer not only efficient, but also secure and trustworthy solutions.

3.2 eSciDoc Solutions

The first eSciDoc solution, Publication Management, will provide basic functionalities and user interfaces for the submission of publication data of multiple types, such as article, conference-paper, poster, report, book etc., along with the metadata needed for proper retrieval and long-term archiving. Structured relations between the publications as well as between affiliated persons and organizations are managed by the solution. The component 'Search & Query' support the discovery of stored publication items, by a user interface as well as via SRU/SRW-interfaces. Full-text search within attached files is provided as well. The component 'Browse and Display' displays the actual items according to the individual sorting and filter mechanisms and provides detailed information for the entry. The structured information on affiliated persons and organizational units allows the user to define views and reports on the data under different aspects. Publication items themselves are persistently identified, (the metadata record as well as each file attached), any individual compilation of publication items, e.g. the MPS yearbook, can be published and persistently identified as well (catalogs). Versioning will be provided for administering the metadata entries as well as for identification of adequate versions of a full-text. The solution will provide a standard simple and complex publication workflow, which can be configured for local needs of internal quality assurance. Various export and import formats are supported to facilitate the integration with individual bibliographic management systems, local databases and external systems. The administration of the system focuses on the needs of different user groups (scientists, local libraries) and provides adequate workspaces and administering functionalities, such as ingestion mechanisms, batch modifications or alert mechanisms.

The second solution developed in the eSciDoc project, the Scholarly Workbench, aims at providing a generic solution for communities in the arts and humanities, to store their digital artifacts and make them “processable” and re-usable within a collaborative environment.

The first basic concepts have been developed in strong cooperation with one partner institute, the MPI for the History of Science and focus on functionalities needed for a collaborative authoring process on digital artifacts, such as images, manuscripts, drawings, transcriptions, annotations. The solution aims at the re-combination of published artifacts, and the integration into other knowledge representations, to support the steadily growing network of digitized cultural heritage. The publication workflow supports a distributed collaborative environment, where users can submit, annotate, and publish collections and bundles of multiple objects, which can be related by structural and semantic relations. As fundamental part of the collaborative environment, users can work offline and modify or extend existing objects in their local environment, followed by the integrated online publication of their result. Annotations and comments to semantic units enrich the published artifacts. The solution supports the versioning of each object as well as of the respective compilations. The persistent identification of relevant published data entities and compilations will be guaranteed. To support the specific requirements for viewing and processing the artifacts in high resolution, specific viewing environments and tools will be developed. The integration of linguistic tools and external services such as language technologies supports the seamless navigation and research process.

4 Technical View on eSciDoc

4.1 eSciDoc as Service-oriented Architecture

The eSciDoc system is designed as a service-oriented architecture (SOA) [7] implementing a scalable, reusable, and extensible service infrastructure. Application- and discipline-specific solutions can then be built on top of this infrastructure. The heterogeneity of the envisioned solutions in addition imposes an efficient handling of different kinds of content.

The service-oriented architecture fosters the reuse of the existing services. An eSciDoc service may be reused by other projects and institutions, either remotely or locally, thus becoming one building block with a broader e-Science infrastructure. At the same time, the SOA approach of eSciDoc comes with other advantages. Instead of a complex and monolithic application, the eSciDoc service infrastructure is rather to be seen as a set of loosely coupled services, which can be specified and implemented independently. This allows for an iterative implementation strategy for services. First services may already be implemented while others are still in their design phase. Based on feedback from early adaptor users (“pilots”), new services can be easily added, thus fulfilling user expectations in a more timely and user-driven manner.

The core technology used to implement the services is based on Java and XML. Instead of building the infrastructure “from a scratch”, the eSciDoc team chose to integrate existing open-source components as much as possible. eSciDoc services in general provide both SOAP and REST [8] style interfaces. This allows for further development of solutions without constraining the selection of the programming languages, thus accelerating their implementation and enabling the involvement of various developer groups. Even simple scripting and “Web 2.0”-style mash-ups are supported.

The eSciDoc service infrastructure groups its services into three service layers: basic services, intermediate services and application services.

The eSciDoc service infrastructure currently does not implement the process layer services. We consider the implementation of process layer services at a later stage. This will enable the orchestration of services.

Basic services provide basic create, retrieve, update and delete (CRUD) operations on data resources such as Items, Containers, Contexts, Organizational Units. Some services in addition provide task-oriented operations that implement part of the core system logic (e.g. changing the status of the resource within the repository). Basic services are stateless and can be used further by other (eSciDoc and non-eSciDoc) service requestors, i.e., clients like services or applications.

Intermediate services represent both a service requestor and a service provider within the eSciDoc SOA. They act as adapters and façades to the basic services or add additional functionality. Intermediate services are stateless and can manipulate their own data in addition. Examples are Item validator, Baskets handling and Duplicate detection service.

Application services compose other services from basic, intermediate and application layer and implement business logic from solution-specific domain. They are candidates for future process-centric services to enable the service orchestration. Examples are Depositing service, Publishing service, Citation manager service, User Management.

Another class of services are "technical" services that enable authentication and authorization. They span vertically all layers of the eSciDoc SOA architecture. Each service operation will invoke the authentication and authorization services.

In addition to services, which come with "machine-friendly" interfaces, eSciDoc provides a set of GUI Tools that offer user-friendly interfaces to business processes. Such tools may be reused within several application solutions and can be further extended to meet domain-specific requirements. GUI Tools thus represent additional building blocks for composite applications development.

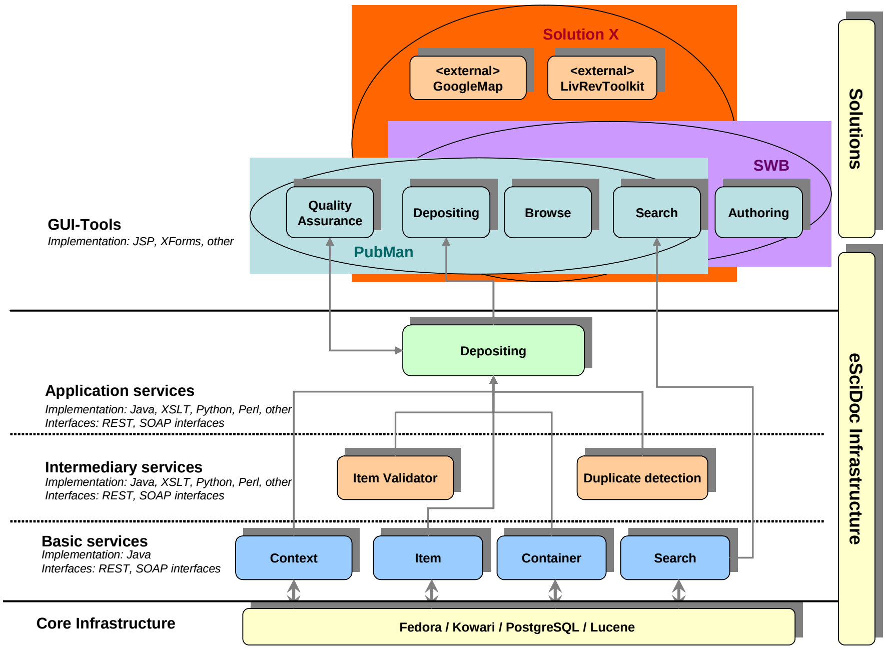

Figure 1: Example of the service stack and related technology stack.

4.2 The eSciDoc Infrastructure

The eSciDoc Infrastructure is a kind of middleware, which encapsulates the repository and implements services of all layers of the service-oriented architecture relevant to the eSciDoc system: the Basic Layer, the Intermediary Layer, and the Application Layer. The eSciDoc Infrastructure is an "enabling technology": Scholars and Scientists can focus on domain-specific application logic when building new applications. It provides them with an existing and proven implementation of common functionality, thus ensuring interoperability and compliance with important standards. Additionally, it allows for the operation of a production environment by a dedicated unit, e.g., a fully-fledged data center. The institutes do not have to care about managing the production services, but can rather concentrate on their scientific and scholarly work.

The Core Infrastructure is mainly built out of existing open-source software packages. Main components are PostgreSQL, JBoss Application Server, and Tomcat Servlet Container. The eSciDoc Content Repository is based on Fedora (Flexible Extensible Digital Object Repository Architecture) [9]. Fedora comes with a Semantic Store (Kowari Triplestore or MPTStore), which allows for the efficient administration of statements about objects and their relations, expressed in RDF (Resource Description Framework) [10]. Related objects form a graph, which can then be queried or used to infer new facts, based on existing RDF.

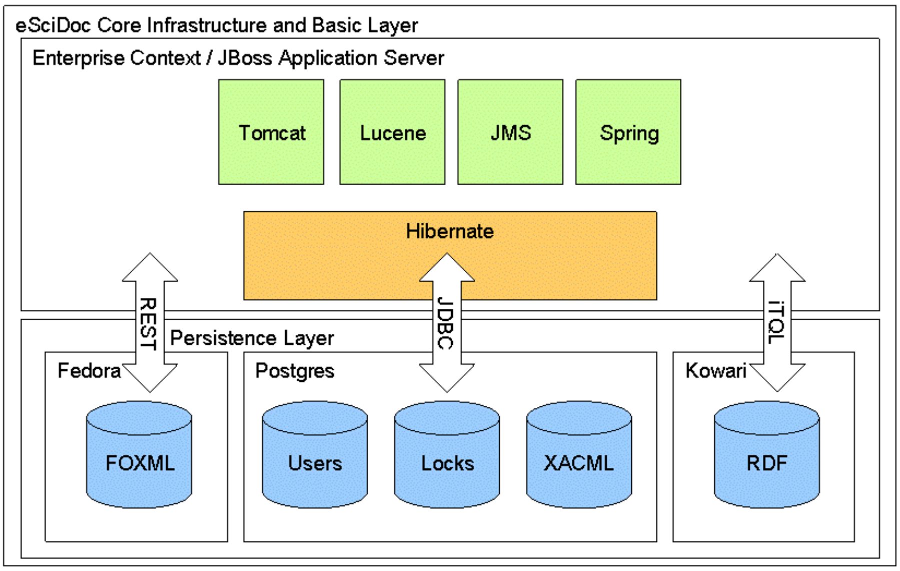

Figure 2: Schematic View of the eSciDoc Core Infrastructure and Basic Services Layer

The eSciDoc Infrastructure is implemented as a Java Enterprise Application (J2EE) [11]. It can be roughly differentiated into the Enterprise Context and a Persistence Layer. The Enterprise Context is deployed to the JBoss Application Server and the Tomcat Servlet Container. The Spring Framework provides a centralized, automated configuration and wiring of the application objects by Dependency Injection and Inversion of Control [12]. The service layer offers web services with REST and SOAP interfaces. The Persistence Layer encompasses specialized solutions for the different types of data: an RDBMS for structured data, Fedora for unstructured data, and Kowari for semantic data.

The Basic Layer implements a set of resource handlers. Each resource handler is responsible for handling of a specific type of a resource. The most important resources are Items, Containers, and Contexts. *Items* are basic objects that represent content entities within the repository, e.g. articles, images, or videos. *Containers* are aggregation objects that allow for arbitrary grouping of items and other containers. Whereas the general layout of Item and Container resources remains the same, they can be further specialized by content types. Content types impose constraints on objects (e.g. allowed metadata schemas, required metadata, allowed file types and mime types for the binary content and specify a set of content type specific properties). *Contexts* represent units of administration for a set of Items and Containers. They are associated with an institutional body responsible for the management of the content.

As already stated before, each resource handler service implements the four basic operations create, retrieve, update, and delete (CRUD). Additionally, filter methods and task-oriented methods (e.g., for changing the status of an object within a content repository) are provided. All services of the basic layer expose both SOAP and REST interfaces. The following example shows how to retrieve the Item resource object with the object identifier `escidoc:123`. Both methods will return the Item resource as XML representation:

– REST: `GET /ir/item/escidoc:123`
– SOAP: `ItemHandlerService.retrieve("escidoc:123")`

The REST API differentiates between resource- and task-oriented methods. All resource-oriented methods use GET, PUT and DELETE verbs. Task-oriented methods use the POST verb, as they are not idempotent.

At present, intermediate and application services implemented in Java invoke the Basic Services via the SOAP API. They use the JiBX data-binding framework [13] to transform the XML representation of resources returned by Basic services, since it allows for the definition of custom mappings to Java classes.

5 Conclusion and Outlook

As the eScience domain is still an evolving field of activities and projects without a well-developed cross-discipline corpus, eSciDoc puts a strong emphasis on being as agnostic as possible to future application scenarios, without losing actual requirements out of focus to avoid silo-like approaches.

The current quarterly release plan reflects this chain of increasing service complexity by first delivering services with well-understood underlying conceptual models. In further releases, more advanced services and solutions will address further requirements. This iterative approach allows for a manageable build-up of complexity. The current implementation plan starts with basic depositing, discovery and retrieval functionalities. Versioning and workflows for data as well as more sophisticated handling of metadata and object relations will follow. This is accompanied by ongoing additions of convenience functionalities to services and solutions, e.g., normative and authoritative data or specific metadata enrichment. Additional discipline-specific solutions are developed to evaluate and improve the concepts and validate decisions, to ensure a pragmatic orientation. By this approach, the eSciDoc infrastructure provides an early basis for organizational integration, while being open for ongoing developments, both by contributors external to the project, and additional future services.

References

1. Website of the eSciDoc Project: <http://www.escidoc-project.de/>
2. Becla1a, J., Hanushevskya, A., Nikolaevb, S., et al. 2006. Designing a Multi-petabyte Database for LSST. *SLAC Publication* 12292. <http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-12292.pdf>
3. Kovac, C. 2003. Computing in the Age of the Genome. *The Computer Journal.* Volume 46, Issue 6, 593-597. <http://dx.doi.org/10.1093/comjnl/46.6.593>
4. Henry, G. 2003. On-line publishing in the 21-st Century: Challenges and Opportunities. *D-Lib Magazine*, Volume 9, Issue 10. <http://dx.doi.org/10.1045/october2003-henry>
5. The ECHO system is currently operated by the Max Planck Institute for the History of Science, see <http://echo.mpiwg-berlin.mpg.de/>
6. See the GAVO project website: <http://www.g-vo.org/>
7. Krafzig, D., Banke K., Slama D. 2004. Enterprise SOA: Service-Oriented Architecture Best Practices. *Prentice Hall.* <http://proquest.safaribooksonline.com/0131465759>
8. Fielding, R.T. 2000. Architectural Styles and the Design of Network-based Software *Architectures*. *Dissertation.* University of California, Irvine. <http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf>
9. Lagoze, C., Payette, S., Shin, E., Wilper, C. 2006. Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries.* Volume 6, Issue 2. Springer Berlin / Heidelberg <http://dx.doi.org/10.1007/s00799-005-0130-3>
10. Manola, F., Miller, E. 2004. RDF Primer. *W3C Recommendation.* <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>
11. See Sun Microsystem's J2EE website: <http://java.sun.com/javaee/>
12. Walls, C., Breidenbach, R. 2005. Spring in Action. *Manning*. ISBN 1932394354
13. See JiBX Framework website: <http://jibx.sourceforge.net/>