DABAR - the national infrastructure for digital repositories

Draženko Celjak[1], Zoran Bekić[1], Marko Cundeković[1], Ljiljana Jertec[1], Miroslav Milinović[1], Alen Zubić[1]

[1] SRCE - University of Zagreb, University Computing Centre, Josipa Marohnića 5, 10000 Zagreb, Croatia, repozitoriji@srce.hr

Keywords
digital repositories, institutional repositories, DABAR, digital asset management, DAM, data curation, open access, interoperability, electronic theses and dissertations, ETDS, Islandora

1. ABSTRACT

Having recognized the need for a national digital asset management infrastructure, SRCE – University of Zagreb University Computing Centre in association with a number of research and higher education institutions in Croatia built a national repository system called DABAR – Digital academic archives and repositories. DABAR provides research and higher education institutions with the common infrastructure necessary to establish a secure, reliable and interoperable institutional repositories in a simple way. Dabar allows institutions to manage their digital assets without the need to deal with technical issues related to the implementation and maintenance of digital repositories.

Until the middle of March 2017, DABAR was used to build two national repositories and 109 repositories of Croatian research and higher education institutions. Currently, it supports curation of electronic theses and dissertations, and papers published in electronic journals. The support for conference papers, book chapters, artistic works (photographs and other audio-visual objects) is being actively developed and a plan to build a support for managing research data sets and educational resources is in place.

The DABAR platform and the community gathered around it are actively promoting Open Access to the results of research and to educational modules and resources.

2. DIGITAL ASSETS MANAGEMENT AND OPEN ACCESS IN RESEARCH AND HIGHER EDUCATION COMMUNITY IN CROATIA

One of the first milestones in promotion and fostering of open access in Croatia was the announcement of the Portal of Croatian scientific journals, HRČAK, in February, 2006. Nowadays, HRČAK hosts more than 420 Croatian journals that publish full text articles in open access. The first time that "the principle of the openness to the public" in Croatia appeared in an official document was in 2006 in the *Science & Technology Policy of the Republic of Croatia* ("Znanstvena i tehnologijska politika Republike Hrvatske 2006. – 2010. godine", 2006). The document stipulated that the research and development results that had been funded by public funds had to be publicly available in open publications or open access databases. A couple of years later, in 2012, *The Croatian Open Access Declaration* was announced and signed by numerous institutions and individuals in Croatia. At that time, only a few institutions had a technical solution for archiving digital assets mainly based on EPrints software. The awareness about the need for the systematic approach to data curation in Croatia was still developing. The first digital repositories in Croatia were HRČAK, FAMENA PhD Collection, FOI digital library, University of Zagreb Medical School Repository, Faculty of Humanities and Social Sciences Institutional Repository and Full-text Institutional Repository of the Ruđer Bošković Institute FULIR. ("OA i OER u Hrvatskoj", 2013)

The importance of Open Access that brought together the scholarly community in Croatia was finally stated in *The Scientific Activity and Higher Education Act* ("Zakon o znanstvenoj djelatnosti i visokom obrazovanju", 2013), which requires universities and higher education institutions to permanently publish theses and doctoral dissertations in public databases (repositories).

Having recognized the Croatian academic and research community's need and aiming to enable it to fulfil its legal obligation, during 2014 SRCE initiated a network of numerous experts and scientists that were interested in digital repositories or had already been working on archiving digital assets for their institutions. Furthermore, SRCE established contact with numerous research and higher education institutions such as the National and University Library in Zagreb, the Ruđer Bošković Institute as well as other higher education and research institute libraries.

As individual institutional solutions for archiving and curation of digital assets were organizationally, financially and technically challenging, SRCE recognized the importance and advantages of building a national e-infrastructure for digital repositories. The vision of a national solution was welcomed not only by institutions that were under a legal obligation to establish digital repositories, but also by other institutions that saw the opportunity to collect and preserve their digital assets in one place.

In addition to the technical challenges in establishing digital repositories, research and higher education institutions were faced with organizational challenges in adopting internal policies and documents that entitle them not only to publish digital material created by their students and staff, but also to publish it according to the principles of open access.

On 4 March 2016, SRCE signed the memorandum of understanding with four institutions within the Croatian academic and research community: the Ruđer Bošković Institute, University of Zagreb School of Medicine, University of Zagreb Faculty of Humanities and Social Sciences and the National and University Library in Zagreb. The institutions agreed to foster organisational, informational and technical development of the national e-infrastructure for digital repositories – DABAR.

3. GOALS OF THE PROJECT

The main goal of the DABAR project was to build a robust and scalable national infrastructure for digital repositories that would enable research and higher education institutions to establish their own reliable and interoperable digital repositories. At the same time, this national infrastructure was expected to promote and follow open access principles.

Furthermore, each institution had to be able to build an institutional digital repository on the institution's internet domain. This was important especially to universities because the metrics that compare universities (e.g. Webometrics[1]) are based on the volume and quality of electronic publications available within universities' internet domain.

The technical requirements for DABAR were to provide a service that is:

- secure and reliable for long term preservation
- interoperable with other national and global infrastructures (e.g. Portal of Scientific Journals of Croatia, OpenAIRE, …)
- scalable (ability to support many institutions and large number and size of digital objects)
- sustainable
- flexible in the sense that it could be customized according to a particular institution's needs.

SRCE, as a major computing centre and the architect of the e-infrastructure for research and higher education community in Croatia, has taken responsibility for the development and maintenance of the necessary infrastructure, which includes: providing the networked computing resources and data storage, taking care of the security and reliability of the whole system and maintaining an application solution linked with appropriate middleware (national identity federation AAI@EduHr).

[1] http://www.webometrics.info/en/Methodology

From the very beginning, SRCE's plans included bringing together the wider community[2] to take an active role in shaping the services. Following tasks were planned:

- setting up working groups from the professional community, primarily advanced users of the system, which would take over the responsibility for:
  - the definition of the information needs of the community
  - the design and maintenance of common metadata profiles and controlled vocabularies
  - participation in education and support for the end users
- raising awareness about the importance of systematic preservation of research and higher education institutions' digital assets
- promotion of open access to research results and educational content.

At the same time, SRCE has aimed to establish its own sustainable and competent team for development, maintenance and continuous improvement of DABAR's services.

4. STATE OF THE DEVELOPMENT AND RESULTS OF THE PROJECT

As a result of the project, a service called DABAR[3] has been developed and released. DABAR is a part of the national infrastructure that enables a simple and free-of-charge set-up of a digital repository for all the institutions in the research and higher education community in Croatia. More information about the architecture and technical aspects of DABAR is provided in section 5.

DABAR's production started on 17 August 2015, and, by 17 March 2017, 109 digital repositories have been requested and set-up in the system. The amount of applications for the set-up of the repository is an indicator that the institutions have recognized the need for systematic management of their digital assets and DABAR as a solution to that need.

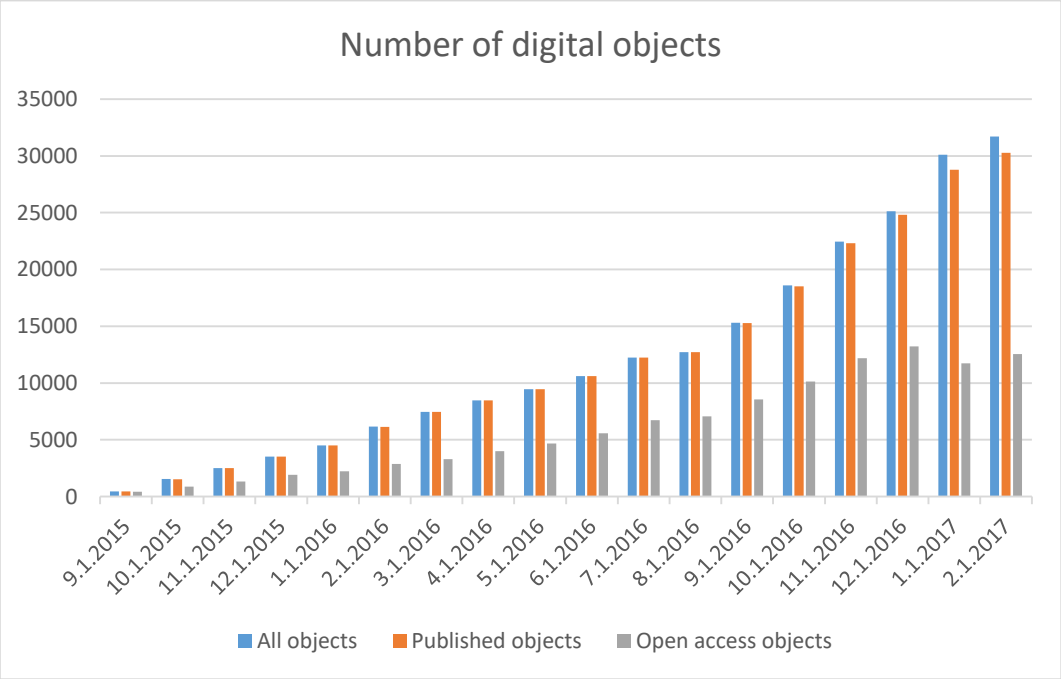

Figure 1. Number of digital objects stored in repositories in DABAR

[2] https://dabar.srce.hr/en/partner-institutions

[3] https://dabar.srce.hr/en

SRCE has been periodically organising educational workshops for repository administrators. By 10 March 2017, a total of 14 workshops with 181 participants were held.

The National and University Library in Zagreb, in collaboration with SRCE and other higher education libraries, has defined metadata descriptions for digital objects "student thesis" and "dissertation and scientific master's thesis". Those definitions were accepted at the national level, which ensures a high degree of interoperability and compliance with international standards. The metadata description for object "paper published in journal" along with needed controlled vocabularies were also defined and agreed upon at the national level.

By 1 March 2017, 30,276 digital objects were stored and described with rich metadata in repositories in DABAR. Figure 1. shows the number of digital objects stored in the repositories in DABAR by months (DABAR statistics, 2017).

In addition to the institutional repositories, through cooperation with the National and University Library in Zagreb, SRCE has set-up two national repositories:

- Croatian Digital Theses Repository at https://zir.nsk.hr
- Croatian Digital Dissertations Repository at https://dr.nsk.hr.

The impact of DABAR is recognized and visible at the national level through strong interest and involvement of research and higher education institutions (National and University Library, universities, colleges, libraries, institutes) resulting in a significant increase in the number of institutions that have decided to establish and maintain their digital repositories. Furthermore, the importance of publication of information in open access has gained momentum. On March 1st, the repositories in DABAR contained 11,744 digital objects available in open access.

During 2016, employees and students of institutions were involved in the process of storing digital objects through the implementation of self-archiving functionality. Eight Croatian universities have set-up their institutional repositories in DABAR. The team that worked on DABAR put a lot of effort in interoperability of DABAR with other repositories and services using the two basic technologies: REST API and OAI PMH.

5. ARCHITECTURE OF DABAR'S INFRASTRUCTURE

DABAR's repository system is based on Islandora[4], an open-source software framework. Islandora is designed to manage and discover digital assets and it has grown to be a reliable system for repository management. Main framework components are Fedora Commons, Apache SOLR and Drupal. Fedora Commons is another open-source project dealing with preserving and managing digital objects while offering exposure of stored resources through RESTful API. The framework uses well-known Apache SOLR for indexing Fedora content for faster metadata access using advanced and optimised search mechanisms. GSearch, as a part of the Fedora package, is used as a bridge for automatic synchronization between SOLR and Fedora content. Drupal is an open-source CMS providing Islandora users front-end for viewing, editing and storing digital objects in Fedora. Islandora's built-in flexibility allows institutions to build large systems that can scale on demand and its modular architecture can be easily used for extending in desired directions.

DABAR ensures constant development and expansion while meeting a demanding functionality roadmap. Research and higher education institutions, repository managers and end users expected rich and accurate metadata and, at the same time, efficiency in storing and describing digital objects. To meet those expectations, it was necessary to rely on existing infrastructures and data sources (Figure 2). One of the first tasks while building DABAR was to enable user authentication through Croatian science and higher education identity federation AAI@EduHr[5] so that the users could use their existing institutional credentials and single sign-on functionality.

The support for depositing electronic theses in repositories in DABAR relies on Information System of Higher Education Institutions (ISVU). ISVU is a national information system which had been, by the

[4] https://islandora.ca/about

[5] http://www.aaiedu.hr/en

end of February 2017, used by 111 Croatian higher education institutions to administer data on teachers, assistants, syllabi, curricula, students enrollments, exams, tuitions, theses etc. In addition, ISVU exposes data through well-documented REST API. In the process of describing and storing theses in repositories, the metadata are fetched from ISVU. Other national registries (for example the Registry of study programs maintained by the Ministry of Science and Education) are used in the process of storing theses and dissertations. The fact that these sources were used for depositing digital objects in repositories and that metadata will be publicly available, *motivated and encouraged* institutions and the Ministry to increase the data quality and accuracy.

While building support for describing and storing papers published in journals to institutional repositories, the development team acknowledged the fact that there were already repositories in which these objects might be described and stored. The HRČAK[6] portal is a platform built for Croatian journal editors for publishing journals in Open Access. On 14 March 2017, it contained more than 155,000 papers available in full text which were described and stored by journal editors of more than 420 journals. To retrieve article metadata and URL of files in PDF format, DABAR uses HRČAK's OAI-PMH[7] interface. For the articles that have a DOI assigned by CrossRef, DABAR uses CrossRef API[8] to retrieve article metadata which was delivered to CrossRef in the process of assigning a DOI to the article. Similar connection is being built for retrieving metadata and PDF's from Croatian Scientific Bibliography (CROSBI) which stores scientific papers with more than 460,000 bibliographic records and more than 30,000 full-text papers available (Hrvatska znanstvena bibliografija, 2017). The plan is to implement metadata retrieval from PubMed Central. PubMed Central is a free full-text archive of biomedical and life sciences journal literature which exposes articles metadata through API[9].

A unique persistent identifier URN:NBN is assigned to each digital object stored in any repository in DABAR. The URN:NBN is assigned by URN:NBN service developed and maintained by the National and University Library in Zagreb.

Besides support for direct input via user interface, digital object can be stored in repositories in DABAR via the DABAR REST API. DABAR REST API was implemented because there were various external sources that already contained the fully described digital objects and had the need to transfer and store the objects in a repository in DABAR. One example of a system that is using DABAR REST API is institutional information system that already has theses stored but doesn't have all the dissemination features of a full-blown digital repository. Another use-case for DABAR REST API is an existing stand alone institutional repository that wants to migrate its content to the repository in DABAR.

An important role of digital repositories is to support dissemination. Besides the user interface, all repositories in DABAR have a built-in OAI-PMH which exposes structured metadata in two standard metadata formats: Dublin Core (DC)[10] and Metadata Object Description Schema (MODS)[11]. All OAI-PMH interfaces of repositories in DABAR are fully compliant with the OpenAIRE Guidelines for Literature Repository Managers 3.0[12] and are ready to be registered as a data provider on the OpenAIRE portal.

The landing pages of the digital objects all have Highwire Press and Dublin Core <meta> tags recommended by Google Scholar Inclusion guidelines for Webmasters[13]. This ensures that digital objects will be included in Google Scholar searches.

---

[6] http://hrcak.srce.hr

[7] https://www.openarchives.org/pmh/

[8] https://github.com/Crossref/rest-api-doc/blob/master/rest_api.md

[9] https://www.ncbi.nlm.nih.gov/home/develop/api.shtml

[10] http://dublincore.org/documents/dces/

[11] http://www.loc.gov/standards/mods/

[12] https://guidelines.openaire.eu/en/latest/literature/index.html

[13] https://scholar.google.com/intl/en/scholar/inclusion.html#indexing

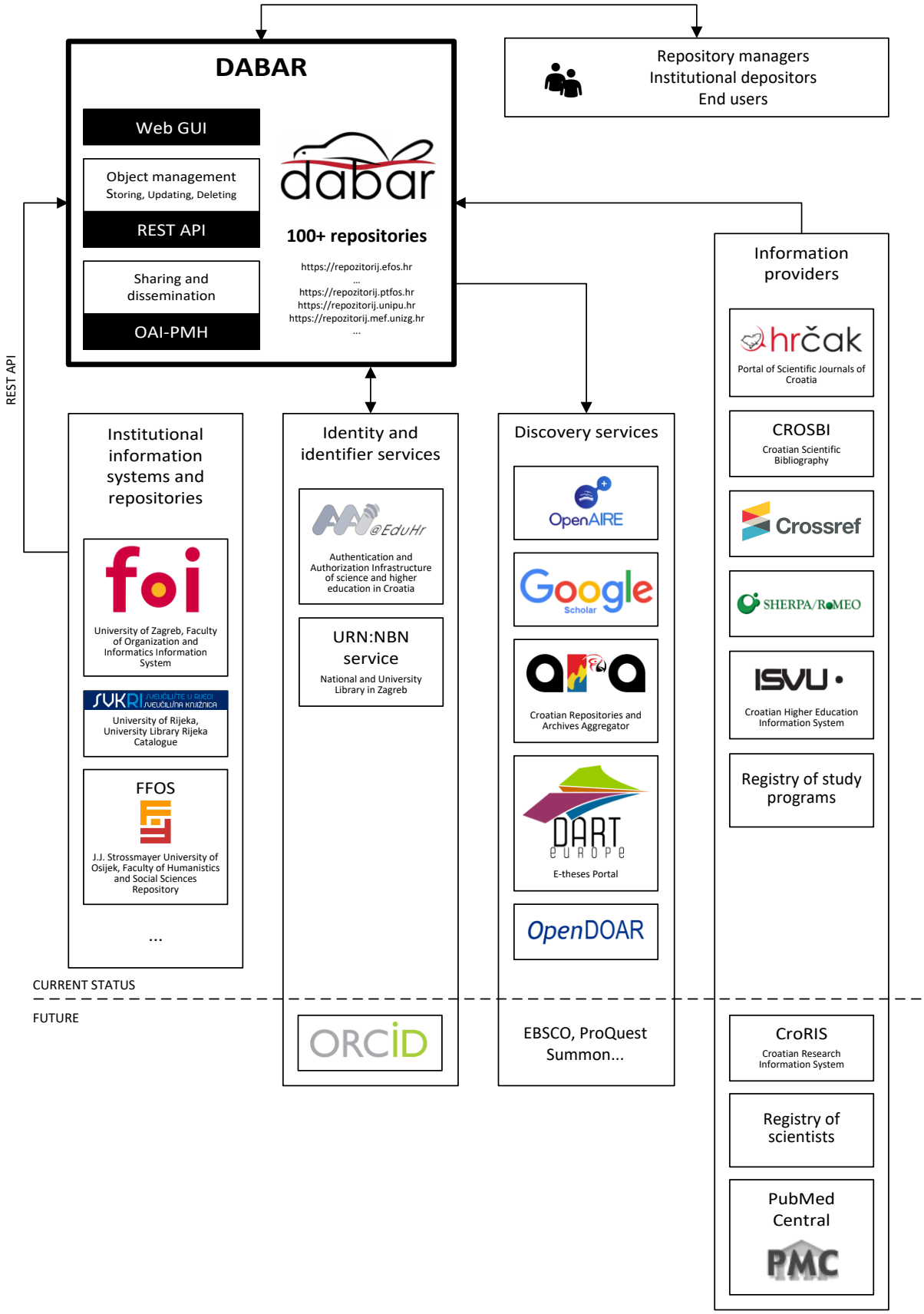

Figure 2. DABAR's integration with national and global e-infrastructure

DABAR's presence on OpenAIRE portal, Google Scholar and other discovery services drives a greater number of readers to repositories and its content and makes the repository content, authors and institutions more visible. Figure 2 ilustrates DABAR's integration with other national and global services.

DABAR's modular architecture is presented in Figure 3. Each of the servers has its unique function in the Islandora framework, extended with continuous integration (CI) server and servers that are mirroring production environment for development purposes. DABAR was built following Software as a Service (SaaS) delivery model in order to provide easy access to a fully featured platform for the institutions which are using it.

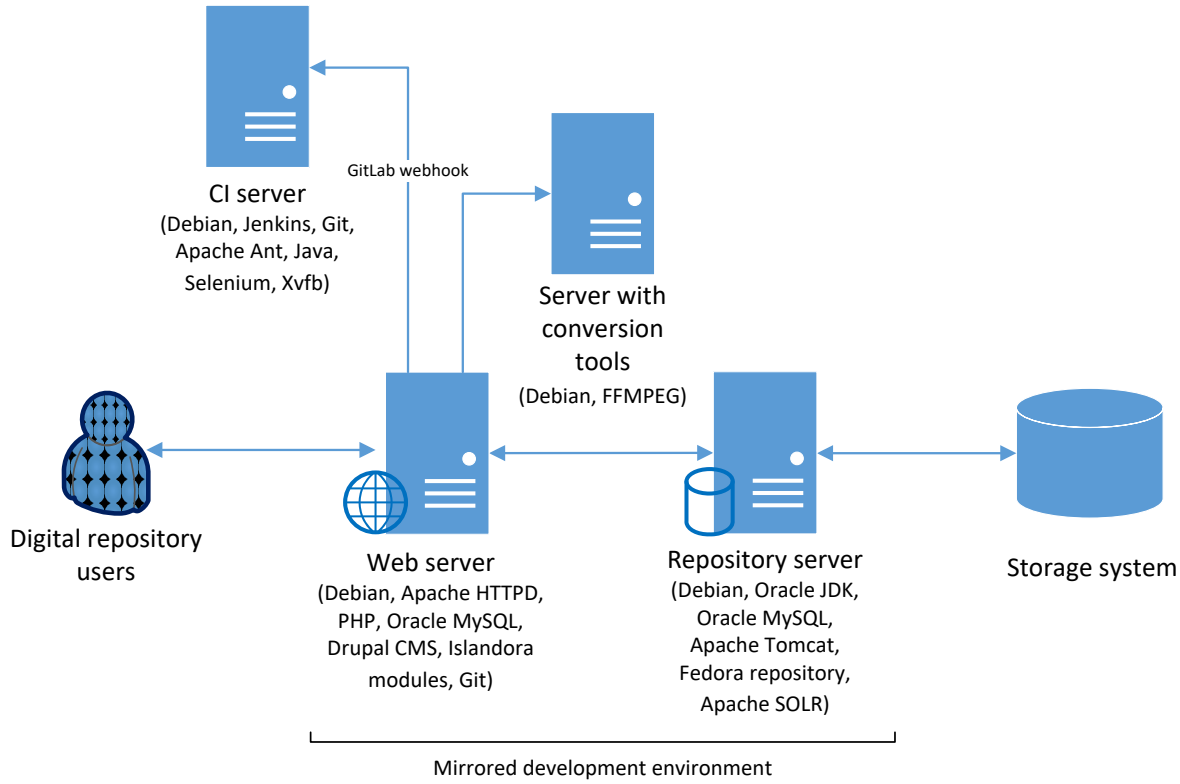

Figure 3. DABAR architecture

6. FURTHER DEVELOPMENTS

DABAR has brought together the research and higher education community providing them with a nationally accepted solution for preservation and dissemination of electronic theses and dissertations, articles published in scientific and professional journals. Still, that is just a good start in fulfilling the institutional needs. Next digital objects that will be introduced during 2017 are conference papers and presentations, book chapters, books and artistic works, including photographs and audio-visual objects. In 2017 we plan to start the work on research data management and educational content.

Given that only 40,8% (11,744 from 28,779) of all available objects in repositories in DABAR are published in open access, a big challenge for DABAR team and community is to promote open access further and support institutions in creating preconditions for publishing in open access.

Interoperability, in particular with national information systems in science and higher education stays in our focus.

7. CONCLUSION

109 digital repositories were established in DABAR's repository system by research and higher education institutions in Croatia. This proves that DABAR has been recognised as a reliable, secure and easy-to-use national infrastructure for the set-up and maintenance of interoperable and sustainable digital repositories. The model of centrally managed national infrastructure helps in reducing the costs of development and maintenance. It provides interoperability for all of its repositories with national and global e-infrastructures. This model is applicable outside the system of science and higher education, for example, in elementary and secondary education system as a repository of educational content or as a commercial platform for hosting digital repositories. It is also applicable in other countries or as an international solution that would meet the need of a specific community.

8. REFERENCES

Croatian scientific bibliography (2017). Retrieved March 15, 2017, from: https://bib.irb.hr/index.html?lang=EN

DABAR statistic: total objects [data set] (2017). *SRCE – University of Zagreb, University Computing Centre*. Retrieved March 15, 2017, from: https://dabar.srce.hr/en/stats

Ministarstvo znanosti, obrazovanja i športa. (2006). *Znanstvena i tehnologijska politika Republike Hrvatske 2006. – 2010. godine*. Retrieved March 13, 2017, from: https://vlada.gov.hr/UserDocsImages/Sjednice/Arhiva/1%20-%20157.pdf

OA i OER u Hrvatskoj (2013). *Sveučilište u Zagrebu, Sveučilišni računski centar Srce website*. Retrieved March 15, 2017, from: http://haw.nsk.hr/arhiva/vol6/472/42691/www.srce.unizg.hr/proizvodi-i-usluge/podatkovne-usluge/otvoreni-pristup-i-otvoreni-obrazovni-sadrzaji/oa-i-oer-u-hrvatskoj/index.html

Zakon o znanstvenoj djelatnosti i visokom obrazovanju (NN 139/13) (2013). *Zakon.hr : Pročišćeni tekstovi zakona*. Retrieved March 17, 2017, from: https://www.zakon.hr/z/320/zakon-o-znanstvenoj-djelatnosti-i-visokom-obrazovanju

9. AUTHOR'S BIOGRAPHIES

Draženko Celjak is head of Data services and collaboration systems at SRCE - University of Zagreb University Computing Centre (Croatia). He studied Information systems at the University of Zagreb, the Faculty of Organization and Informatics Varaždin. Over the course of his professional career, he worked as a project leader and/or developer on different national systems dealing with digital repositories and archives: DABAR - Digital academic archives and repositories, HRČAK – Portal of scientific journals of Croatia, Croatian Web Archive (HAW), ARA - Aggregator of Croatian Repositories and Archives. In 2015 he received the "Tibor Tóth" award from Croatian Information and Documentation Society (HID) for a significant contribution in the field of information science. His main areas of professional interest are the use of web technology, web archiving and indexing, semantic web, linked data and digital repositories. (Linkedin: https://www.linkedin.com/in/draženko-celjak-18249893/)

Zoran Bekić is the director/CEO of SRCE - University of Zagreb University Computing Centre (Croatia). He received a BSc and MSc degrees in applied mathematics and cybernetics from Lomonosov Moscow State University and a PhD in Information science from the University of Zagreb. He has coordinated or participated in a number of projects focused on development and improvement of components and services of national research and educational e-infrastructure, including fields like information systems, advanced computing, e-learning, data curation, open access and open educational resources. He has authored more than 70 research or technical papers. (LinkedIn: https://www.linkedin.com/in/zoranbekic/)

Marko Cundeković is a computer specialist at SRCE - University of Zagreb, University Computing Centre (Croatia) in Data services and collaboration systems sector. The main project he is involved in is DABAR - Digital academic archives and repositories. His fields of interest are linked data, semantic web, big data and artificial intelligence. He studied Computer Aided Engineering at the Faculty of Mechanical Engineering and Naval Architecture at the University of Zagreb.

Ljiljana Jertec is a librarian and computer specialist at SRCE – University of Zagreb University Computing Centre (Croatia). She has a Master's degree in Library and Information Science which she studied at the University of Zagreb, the Faculty of Humanities and Social Sciences. She works in SRCE's library and on the following projects: HRČAK – portal of scientific journals of Croatia and DABAR – Digital academic archives and repositories. Her main interests are digital repositories, metadata, metadata schemes, scholarly communication and open access initiatives.

Miroslav Milinović is the Assistant Director for Information and Middleware Infrastructure at SRCE – University of Zagreb University Computing Centre (Croatia). He has been leading SRCE's research in the areas of middleware, web technologies and digital archives for over 15 years. He was leading the development of HRČAK – Portal of scientific journals of Croatia, Croatian Web Archive (HAW) and Croatian identity federation AAI@EduHr. He received a MSc degree in computer science from the University of Zagreb, the Faculty of Electrical Engineering and Computing. His main professional interests are related to middleware, metadata, web technologies and digital repositories.

Alen Zubić is a computer specialist at SRCE - University of Zagreb, University Computing Centre (Croatia) in Data services and collaboration systems department. His fields of interest are machine learning, neural networks, digital infrastructure engineering and software development process management. He studied Engineering design specializing in Design and Product Development at the University of Zagreb, the Faculty of Mechanical Engineering and Naval Architecture.