The e!DAL Java API: Sharing and Citing Research Data in Life Sciences

Daniel Arend*, Jinbo Chen, Christian Colmsee, Steffen Flemming, Uwe Scholz and Matthias Lange

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, OT Gatersleben, 06466 Stadt Seeland, Germany

{*corresponding author: arendd@ipk-gatersleben.de}

1 Background

The data life cycle from experiments to scientific publications follows in general the schema: experiments, data analysis, interpretation, and publication of scientific paper. Besides the publication of scientific findings, it is important to keep the data investment and ensure its future processing. This implies a guarantee for a long-term preservation and preventing of data loss. Condensed and enriched with metadata, primary data would be a more valuable resource than the re-extraction from articles. In this context, it becomes essential to change the handling and the acceptance of primary data within the scientific community. Data and publications should be honored with a high attention and reputation for data publishers.

2 The e!DAL data publication pipeline

Here, we present new features of the e!DAL Java API (http://edal.ipk-gatersleben.de) as a lightweight software framework for publishing and sharing of research data [1]. e!DAL stands for electronic Data Archive Library. Its main features (Table 1) are version tracking, management of metadata, information retrieval, registration of persistent identifier, embedded HTTP(S) server for public data access, access as network file system, and a scalable storage backend. e!DAL is available as an open- source API for a local non-shared storage and remote usage to feature distributed applications.

adfa, p. 1, 2011.

Features	**Applied frameworks, standards and APIs**
Version Management – sequence of versions for the data set and its metadata	• H2 database • Hibernate • File system
Metadata Management – minimal set of technical and administrative metadata to ensure the mid-term data access of the stored data set	• DublinCore
Information Retrieval – search and retrieve relevant data sets for keyword queries over the metadata	• Apache Lucene • Hibernate Search • Apache SolR
Persistent Identifiers – provide persistent identifiers for an long-term stable public access of published data sets	• DOIs • URLs
Data Security – fine grained authorization of API methods, referred object and authenticated subjects (user)	• AspectJ • Java Authentications & Authorization API
Interoperability – seamless integration into existing infrastructures/tools	• local/remote Java API • GUI components • WebDAV • HTTP(S) Server

Table 1. Main features of e!DAL are listed in column one. For its implementation used frameworks, standards and APIs are referenced in column two.

The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) is an approved data center belonging to the international DataCite consortium (http://www.datacite.org/) and applies e!DAL as data submission and registration system. In the latest version the focus was to extend the features for the registration of Digital Object Identifier (DOI) and the development of a simple, but sufficient approval process to regulate the assignment of persistent identifier. An intuitive publication tool (Figure 1), allows uploading your data into your own private repository over the web and getting a DOI to permanently reference the datasets and increase your "data citation" index.

The e!DAL approval process for data publication is based on an email notification system. After successfully proven a request for a DOI, the API automatically transfer

all necessary files and metadata to DataCite and reply an email to the submitting user with his final assigned DOI. This ID permanently references to a virtual content page.

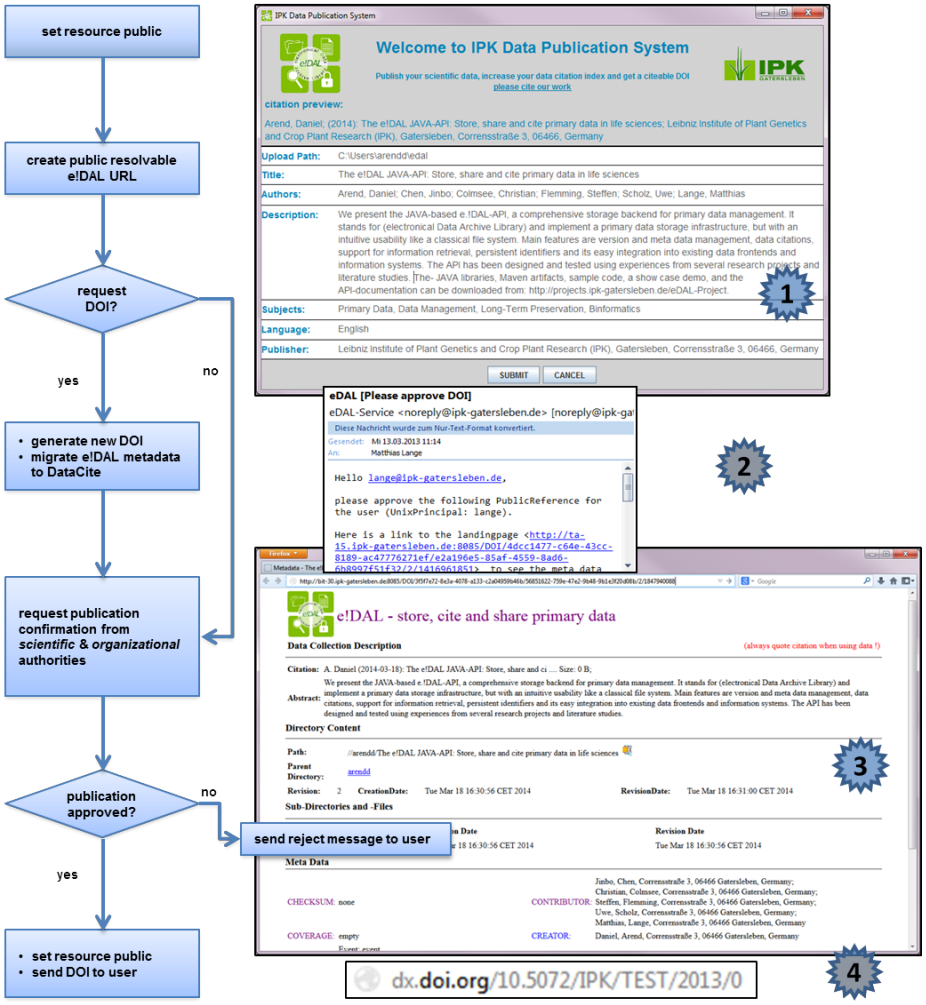

Fig. 1. Schema of the e!DAL Publicaton workflow. e!DAL provides an easy usable publication interface for DOIs with a graphical user interface (1) and a simple approval process, which allows to define different reviewers, who can accept or revise every requested ID. The process is based on an email notification system (2). When the survey was successful, the API automatically transfer all necessary files and metadata to DataCite and the requesting user get an email with his final assigned DOI (4). This ID permanently references to a virtual content page (3), where you can download the files and check the corresponding metadata.

In addition we implement some new graphical components, like an easy installation/demo wizard, to simplify the deployment of a repositories using e!DAL.

3 Conclusion

e!DAL is a lightweight software framework for the management, publication, and sharing of research data. It is designed to turn sets of primary data into citable data publications. This is particularly important for the life sciences, where there is a big gap between the rate of data collection and the rate of data publication. Its well-defined API supports seamless integration into existing data-management software and infrastructures. In addition, e!DAL can be used as a supplement to manage primary data. Furthermore, its modular architecture and incorporated standards ensure version management, documentation, information retrieval, persistent identification, data security, and data publication. Developed within a context for the life sciences, e!DAL has many generic features that make it easily and readily applicable to other areas of science faced with similar needs. The e!DAL software is proven and has been deployed into the Maven Central Repository. Documentation and Software are also available at: http://edal.ipk-gatersleben.de.

4 References

[1] - D. Arend, M. Lange, C. Colmsee, S. Flemming, J. Chen and U. Scholz. *The e!DAL JAVA-API: Store, share and cite primary data in life sciences*. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2012, pages 1–5. 2012. DOI: 10.1109/BIBM.2012.6392737