Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 13-18.
ISBN 978-88-8303-295-0. EUT, 2010.

An online authoring and publishing platform for field guides and identification tools

Gregor Hagedorn, Gisela Weber, Andreas Plank, Mircea Giurgiu, Andrei Homodi, Cornelia Veja, Gerd Schmidt, Pencho Mihnev, Manol Roujinov, Dagmar Triebel, Robert A. Morris, Bernhard Zelazny, Edwin van Spronsen, Peter Schalk, Christian Kittl, Robert Brandner, Stefano Martellos, Pier Luigi Nimis

Abstract — Various implementation approaches are available for digital field guides and identification tools that are created for the web and mobile devices. The architecture of the "biowikifarm" publishing platform and some technical and social advantages of a document- and author-centric approach based on the MediaWiki open source software over custom-developed, database driven software are presented.

Index Terms — field guides, flora, fauna, identification tools, social software, DELTA, SDD, MediaWiki, agile development.

1 Introduction

Digital identification tools may be simple picture guides, printable tabular tools, or interactive tools (single-access, multi-entry, or multi-access keys). A mixture of tools and richly illustrated species pages or glossary definitions is often required. The EU-funded *KeyToNature* project provides a wide spectrum of such tools: together with the "biowikifarm.net", it integrates both the tools and their content. We describe here the architecture and components or this internet-based collaborative authoring and publishing platform.

G. Hagedorn, G. Weber, A. Plank are with the Julius Kühn-Institute, Federal Research Centre for Cultivated Plants, Inst. for Epidemiology and Pathogen Diagnostics, Königin-Luise-Str. 19, D-14195 Berlin, E-mail: gregor.hagedorn@jki.bund.de – M. Giurgiu, A. Homodi, C. Veja are with the Telecomm. Dep., Technical Univ. of Cluj-Napoca, Cluj 400027, Romania – G. Schmidt is with the Institut f. Lern-Innovation, Univ. Erlangen-Nürnberg, D-91052 Erlangen – P. Mihnev, M. Roujinov are with BIKAM Ltd., Sofia 1505, Bulgaria – D. Triebel is with the Center of the Bavarian Natural History Collections, Menzinger Str. 67, D-80638 Munich – R. A. Morris is with the Univ. of Massachusetts, USA – B. Zelazny is with the Internat. Soc. for Pest Information (ISPI) – E. v. Spronsen and P. Schalk are with ETI Bioinformatics, Amsterdam, The Netherlands – C. Kittl, R. Brandner are with evolaris next level GmbH, A-8010 Graz – S. Martellos and P. L. Nimis are with the Department of Life Sciences, Univ. Trieste, I-34127.

2 The MediaWiki Software Architecture

The architecture of the biowikifarm publishing platform is based on the "MediaWiki" open source authoring system [1] that is also used by projects of the Wikimedia Foundation (e. g., the Wikipedias, Wikispecies, Wikisource, or the Commons Media Repository [2]). MediaWiki provides an object oriented document storage model of medium granularity (titled chapters called "pages", rather than whole works). The storage model is akin in many aspects to the currently developed "nosql" database management systems [3] (predating these developments, however, MediaWiki typically uses mysql). Namespaces provided by the storage model allow to re-use the basic model for 1st-class content objects as well as for building objects used in hypertext inclusion. Examples for the latter are media items (binary plus metadata) in the "File" namespace or programming blocks and rich text fragments in the "Template" namespace [4].

The template model provides for flexible schema development. Each template defines a class with freely definable attributes (equivalent to an "entity type"), instances of which can be freely embedded into other objects. Template instances can be hierarchically nested.

The MediaWiki platform is a strong open content and social networking platform. Essential features are the support of the requirements of creative commons licenses (perpetuating licences, tracking contributions and attributing all authors of text and media), a version management and comparison system making changes in a large community transparent to the end user, and a layered development system empowering the community to participate in the functional development of the system.

The latter aspect helps to overcome the discrepancy between user needs and developer actions. Traditional software development requires cycles of planning, use-case and information modelling, piloting, implementation, testing, and rollout, often resulting in slow and inflexible development. Although MediaWiki uses an agile variant of this cycle (involving continuous code integration and live alpha version testing), the php-based core code still suffers from slow development. However, the domain of slow development has been minimized. An event driven extension system provides for an ecosystem of independently developed and tested php-based extensions. Furthermore and highly relevant to the success of MediaWiki projects, the domains limited to developers and server owners are supplemented by further layers (templates, CSS, and JavaScript) that are under the control of the content-editing community:

The templating system enables authors to define and render their own data storage and functional schemata. An unlimited number of templates can be defined, and instances conforming to these schemata (typed and semantically defined fields) can then be inserted in many content objects. Templates are central to the ability of MediaWiki to empower the experts in a given knowledge domain to experiment and achieve information schemata satisfying to their needs. For example, *KeyToNature* defined schemata for media metadata and identification keys. The functionality of templates is limited to prevent detrimental influence on the server, limiting possible malfunctions to those objects that include them.

As a negative point, the templating language has arisen as a unique ad-hoc development, may be difficult to learn, and has no debugging support. Interestingly, this may be a result of social engineering to limit the number of users creating new templates on Wikipedia.

Further layers are the CSS and JavaScript integration. Like templates, these layers are stored as normal MediaWiki objects, profiting from the version control and comparison functionality. Since CSS and JavaScript involve potential security concerns, editing of these layers is limited to content administrators. The community focus of these layers was very positive in the *KeyToNature* project and supported multiple more or less successful approaches to field guides and identification tools.

3 The Biowikifarm

The virtual server is designed as a multi-project platform, enabling the joint administration of a large number of separate wikis. Each wiki can be maintained under its own domain name (owned by partners). Whereas the content administration of each wiki is independent, significant synergies are created by managing multiple MediaWikis on a single "wiki farm".

Presently, the biowikifarm hosts the main *KeyToNature* portal, national *KeyToNature* portals (*pedagogical handbook*, *Offene Naturführer*), the *International Society for Pest Information Wiki*, *LIAS glossary*, *Diversity Workbench*, and the *Deutsche Phytomedizinische Gesellschaft Wiki*.

4 Platform Customization Components

4.1 Media management

The biowikifarm maintains two local media repositories for sharing media between all wikis on the platform. The "OpenMedia" repository is the primary repository for Creative Commons-licensed media. It is supplemented by a "SpecialMedia" repository for media that cannot be openly licensed and are available only under bilateral agreements.

Furthermore, the "Commons" repository with over 7 million images is directly integrated through a web service API. All items from Commons are directly usable as if they were available locally. One problem initially encountered was that the Commons servers may occasionally drop web service requests if overloaded. This could be solved by implementing a license-compatible delayed caching solution (every 10 min. in background).

MediaWiki guarantees the attribution requirements of most Creative Commons licenses by linking media usage to a metadata page containing creators and license information. This page also shows images in a higher resolution. However, displaying this information forces the user to navigate away from the present page. Our own usability studies have shown that users expect an enlarged version of the image without leaving the page context and are confused by the default functionality. A JavaScript based image zooming facility was

therefore added to biowikifarm. The first click on an image will enlarge it in an overlay to the page context, to the maximum extent supported by source image and device resolution. The licensing requirements are fulfilled by presenting a link to creator, copyright, and license information as part of this overlay.

4.2 Metadata and information management

Metadata stored in MediaWiki templates are supported by a customized method. A MediaWiki extension harvests all first level templates (on Wiki pages or inside text files uploaded as attachments) and stores the field-value pairs in a MySQL database for fast access. A web service then provides for queries or recent changes, exposing the data as xml for downstream processing (Fedora Commons, GSearch).

By enabling Semantic MediaWiki (SMW), syntactically defined template schemata are semantically annotated using standard ontologies (Dublin Core, FOAF, SIOC). This allows direct semantic metadata search and inference as well as exposure in the OWL/RDF format. Semantic queries can be embedded, creating dynamical content in wiki pages (outside of metadata, SMW is presently further extensively tested by ISPI).

An embedded Flash application, the MediaIBIS search tool, searches the metadata objects stored in the *KeyToNature* online repository. It has a user-friendly multilingual interface and supports both simple and advanced queries. Details are presented in a separate contribution [5].

4.3 Embeddable identification tools

The platform provides several embeddable identification tools. DELTA datasets can be embedded through NaviKey [6], SDD data through IBIS-ID [7] or Xper^2 [8]. The embedding is achieved through a simple custom “IdentificationTool” extension. To embed a tool, users only need to write a simple statement like: *<IdentificationTool>tool=NaviKey5 config=… NavikeyConfig.xml</IdentificationTool>*. DELTA and SDD data may be placed on wiki pages (rather than in binary attachments) and remain directly editable in plain text mode.

4.4 Wiki-editable single-access identification tools

Single-access keys (i.e. tools with a fixed dicho-/polytomous structure) are implemented in a more direct manner than the multi-access keys. They are based on a combination of templates, CSS and JavaScript, all of which are directly editable by administrators of the wiki (no intervention of server administrators is necessary). On any given wiki page one or several single-access keys can be freely embedded as a structural element in a rich-text layout. The keys provide both a tabular, printable view and an interactive (step-by-step) mode [9]. Details about the wiki key implementation are presented in this volume [10].

5 Sustainability and Scalability

Maximizing sustainability in the face of continuous hardware- and software evolution was a major design priority. Hardware independence can be relatively easily achieved by means of server virtualization, making entire servers easily portable from one physical machine to another. Service is assured by follow-up projects (until 2013) plus a longer-term maintenance pledge of the SNSB IT Center (the SNSB is the government agency for the natural history collections of Bavaria).

Software sustainability is more difficult to achieve. The model of isolated systems maintained in stasis for long periods is not applicable to web software that is dependent on a complex software environment and under permanent threat of malicious attacks. Whereas major publishers achieve permanent redevelopment for their in-house-developments, even mid-sized publishers and software developers have often failed to find the necessary resources. Perhaps the majority of internet offers in biodiversity that were backed by scientific institutions or individuals have therefore ceased to exist. A possible solution is built on three pillars: a) building on a carefully chosen open source software that is supported by a large community with a long-term perspective; b) minimizing project-specific custom developments and partitioning them into small, well documented modules (reducing complexity and the steepness of the learning curve for new developers); c) building the platform to the needs of multiple projects, aggregating available resources and achieving synergies.

We consider the long-term sustainability perspective of MediaWiki to be optimal. It is actively developed, the content of the Wikimedia foundation projects tied to the software makes it highly unlikely that it is abandoned in favour of another project, and version upgrades are always fully automatic (in contrast to some other content management systems that require considerable resources to move from one version to another).

Our own developments are designed to be as modular and layered as possible. They involve small php extensions, a set of templates that can be maintained independent of newer developments, and CSS and JavaScript development. Except for the php extensions, the components are directly editable over the web and can be maintained by a community of users and developers.

An attractive feature of the combination of templates and JavaScripts is their locality to specific documents. The system offers the option to run older identification tools in parallel with newer developments. While this may lessen user experience uniformity, it reduces the analysis and testing requirements for new ideas, enabling agile developments in the future.

Finally, and of great importance to scientific publishing, the principle of locality also applies to content. Scientific knowledge is a stage in a development, no final truth. Opinion may often (yet) matter. Unlike typical databases, the platform assumes no homogenous single truth. Dissenting opinion may be published and outdated knowledge may be retained (adding pointers to updates, etc.). Conventional databases may support dissent (e. g., alternative taxonomic hierarchies), but these expensive solutions are typically limited to a specific aspect. On a wiki platform, any update requires no analysis whether it would

corrupt relational assumptions of older publications – contributing greatly to scalability and sustainability.

6 Conclusion

The MediaWiki-based platform is suitable for the development of collaboratively edited flora and fauna projects. It is powerful, extensible and long-term sustainable. We have successfully implemented a set of native or embedded components. Molecular identification extensions are, however, yet missing. The present platform can be adapted to other purposes in order to create an open source online community of such tools and the scientific interests around them. We welcome further partners to share the platform's use and management.

Acknowledgement

This work was supported by the *KeyToNature* Project, ECP-2006-EDU-410019, in the *e*Content*plus* Programme.

References

[1] MediaWiki software, http://www.mediawiki.org/wiki/MediaWiki, 2010-07.
[2] Wikimedia Foundation Projects: http://wikimediafoundation.org/wiki/Our_projects 2010-07.
[3] MediaWiki Templates, http://www.mediawiki.org/wiki/Templates, 2010-07.
[4] NoSQL databases (overview). http://nosql-database.org/, 2010-07.
[5] M. Giurgiu, A. Homodi, C. Veja, G. Hagedorn and P. L. Nimis, "A search tool for the digital biodiversity resources of *KeyToNature"*. In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 19-24, 2010.
[6] D. Neubacher, and G. Rambold. NaviKey, a Java applet and application for accessing descriptive data coded in DELTA format. http://www.navikey.net. 2005 (onwards), 2010-07.
[7] M. Giurgiu, G. Hagedorn, and A. Homodi, "IBIS-ID, an Adobe FLEX based identification tool for SDD-encoded multi-access keys". *Proc. of TDWG 2009*, 9-13 Nov. 2009, Montpellier, p. 90, 2009.
[8] V. Ung, G. Dubus, R. Zaragüeta-Bagils and R. Vignes Lebbe, "Xper^2: introducing e-taxonomy". *Bioinformatics,* vol. 26 (5), pp. 703-704; see also http://lis-upmc.snv.jussieu.fr/lis/?q=en/resources/software/xper2, 2010.
[9] S. Opitz and G. Hagedorn, "The jKey wiki key player and builder2. *Proc. of TDWG 2009*, 9-13 Nov. 2009, Montpellier, 2009.
[10] G. Hagedorn, B. Press, S. Hetzner, A. Plank, G. Weber, S. von Mering, S. Martellos and P. L. Nimis, "A MediaWiki implementation of single-access keys". In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 77-82, 2010.