medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin

Krzysztof Nowak, Bruno Bon

medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin

Krzysztof Nowak[1], Bruno Bon[2]

[1] Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland

[2] Institut de recherche et d'histoire des textes, CNRS, Paris, France

E-mail: krzysztofn@ijp-pan.krakow.pl, bruno.bon@irht.cnrs.fr

Abstract

medialatinitas.eu is a lightweight Web application which integrates dictionaries, corpora and encyclopaedic resources for Latin. The integration takes places principally on the level of the user-friendly interface, so no explicit links between resources are provided. The main objectives of *medialatinitas.eu* are: improving access to distributed data; challenging separation of linguistic and encyclopaedic information in lexicographic description; compensating for deficiencies of existing lexicographic resources; building community of users who apply computational methods in their study of Latin texts.
As for the architecture, *medialatinitas.eu* is implemented as as mashup application: user's query (as of now, only lemma search is supported) is processed and despatched to both local and distant services (RESTful APIs, SPARQL endpoints); the results are subsequently returned and displayed on the main page as a set of separate widgets. The widgets may contain short concordance lines and tables, but special attention has been given to alternative ways of content presentation, namely charts and visualisations. The widgets are provided with rich graphical hints and hold together thanks to such narrative devices as interpretative notes or explicative commentary. As a whole the widgets contribute to extensive description of Latin lemmas according to their grammatical, semantic and cultural properties.

Keywords: lexicographic mashup, data reuse and integration, visualisation, dictionary-corpus interface, Medieval Latin

1. Introduction

Latin was one of the most widely used languages in European history. In its spoken and written form it was the language of daily communication, law, literature, and science for over fifteen centuries on the territory stretched from Spain to Germany to Poland and from Sweden to Croatia to Italy. This geographical, chronological and functional variation is reflected in a large number of texts which, in turn, gave rise to a vast body of secondary literature of which dictionaries form an essential part.

The multifarious resources, even if partly digitised by now, remain still widely dispersed and do not easily lend themselves to integrated search. Moreover, separate electronic text collections usually cover only small proportion of the texts preserved to our times and do not have any pretensions to representativeness.

Often, they would also be available only through an interface that does not allow for any subtler query. As for the electronic dictionaries, their selective spatio-temporal coverage, multilingual definitions, and differing editorial styles make that they cannot be said to account for Latin development in any systematic way if consulted separately.

medialatinitas.eu is a web application which aims at meaningful integration of textual, lexicographic and encyclopaedic resources for Latin through a user-friendly and attractive interface. It is also an attempt to generate a coherent narrative from incomplete data despite variety of technologies in use. The integration is said to be shallow, since the heterogenous content (dictionaries, encyclopedias and corpora) has been linked only to the degree needed for its unified query and retrieval. It takes place, then, at the level of the web interface which, thus, constitutes presentational layer and a point of access to the services running in the background. At the moment, *medialatinitas.eu* is intended in particular for academic audience (lexicographers, linguists, historians etc.), but teachers and students of the medieval literature should find it useful as well.

2. Data, Goals, Design

2.1. What to integrate: data

medialatinitas.eu makes extensive use of the existing digital resources for Latin language and culture. The data which are going to be integrated within the web application may be roughly divided into three groups (Figure 1):

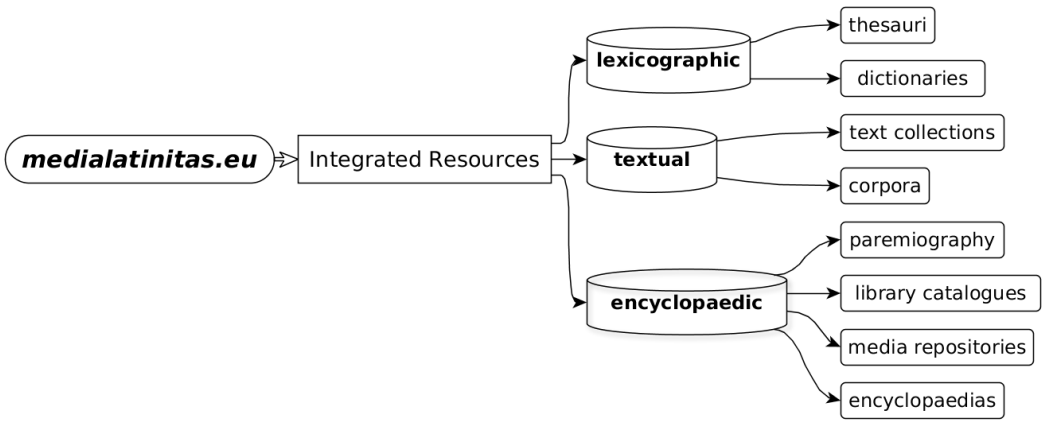

Figure 1: *medialatinitas.eu* resources: general outlook.

1) lexicographic resources: dictionaries of Classical, Medieval and Modern Latin, both academic (e.g. *Novum Glossarium Mediae Latinitatis*, *Lexicon Mediae et Infimae Latinitatis Polonorum*) and community-based (e.g. *Latin Wiktionary*); dictionaries and thesauri of ancient and medieval placenames, gazetteers (*Pelagios Project*, *Orbis Latinus*, *Getty Thesaurus of Geographic Names*, *GeoNames*);

2) corpora (e.g. *Fontes. Corpus of Polish Medieval Latin*, *Croatiae Auctores Latini*)

and text collections (e.g. *Perseus Project*, *Patrologia Latina* etc.);

3) encyclopaedic resources: encyclopedias (*Wikipedia*, in particular its Latin version), paremiological resources (*Latin Wiktionary*), document and image repositories (*Europeana*), library catalogues (*Internet Archive*, *Open Library*), lists of medieval authors (e.g. *Novum Glossarium*, *VIAF*), hybrid resources (e.g. *BabelNet*).

Regarding their origin, the vast majority of resources was created by external institutions and only very few are the result of in-house projects (*Novum Glossarium*, *LMILP*, *Fontes*). As one may suppose, this and the format the data come imply different strategies of access and reuse, and contribute to the complexity of the integration task, as the resources are mostly exploited „as they are".[1] In-house dictionaries come originally as TEI-conformant XML files based on a shared encoding scheme. Both external and in-house corpora were delivered as XML files containing lightweight document mark-up for meta-data and structural features of the text. Each corpus text was tokenised and annotated with PoS and lemma labels. The annotation was performed using the *TreeTagger* (Schmid 1994). The Latin parameter file that the tagger requires was based mostly on the texts from the *Perseus Digital Library* and the *Index Thomisticus*; however, this is likely to be changed in the nearest future, once the work on the Medieval Latin parameter file comes to an end (*Omnia Project TreeTagger*).

The majority of external resources is exploited through their public RESTful APIs or SPARQL endpoints, so the *medialatinitas.eu* remains to some degree agnostic of the original data formats or encoding schemes (Figure 2). Regardless of their origin, even the locally hosted data are exposed to the web application through the APIs:

- dictionaries deployed in an *eXist-db* instance are exposed through respective RESTful API;

- textual corpora are deployed in a *CQPWeb* (Hardie 2012) instance; since for the moment *CQPWeb* does not offer a web API, it is used only as an advanced corpus research tool;

- OCR texts and less-structured text collections are stored in *eXist-db* Lucene-based indexes and exposed through a RESTful API.

1For explanation, see below.

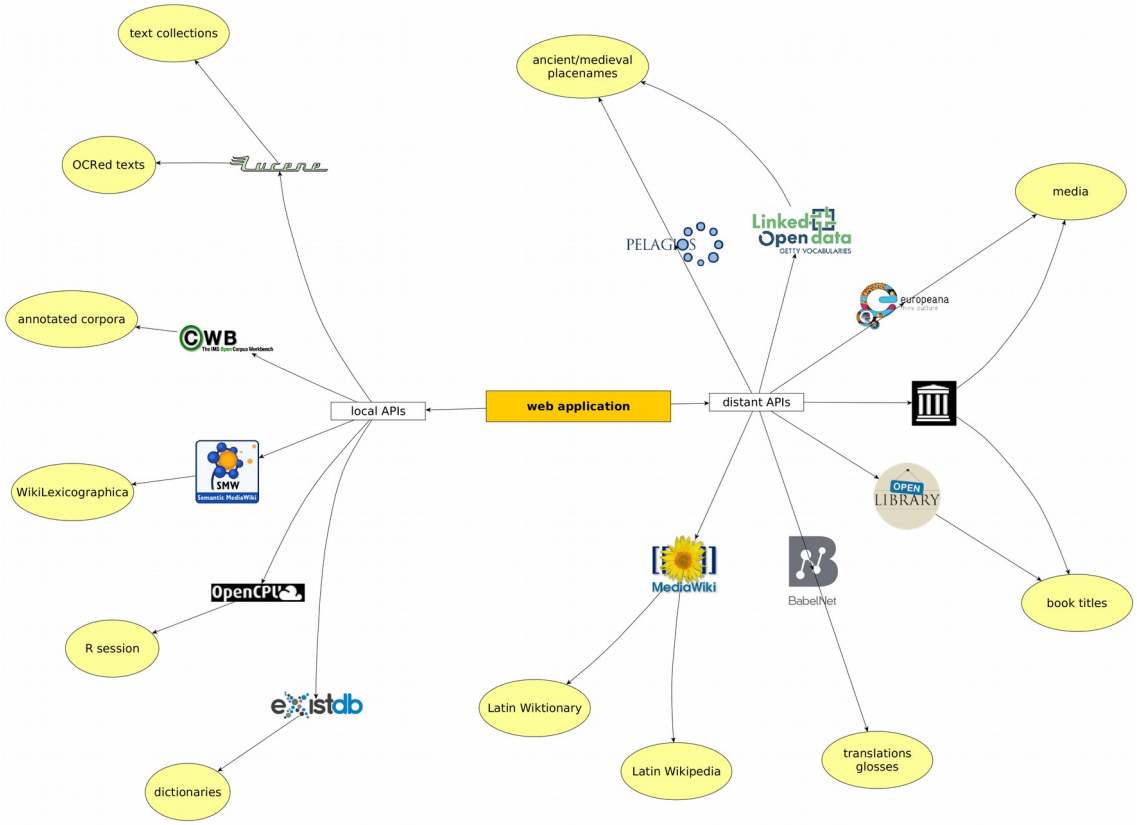

Figure 2: *medialatinitas.eu*: exploited APIs.

Yet, the role of locally running services is by no means limited to only exposing data, since they also serve to enrich, compute on and prepare data for subsequent display:

- the *WikiLexicographica* (Bon & Nowak 2013), an implementation of the *Semantic MediaWiki* (Krötzsch et al. 2006), combines dictionaries with geographical and chronological dimension, thus enabling rich data representation;

- an *R* (R Core Team 2015) session is exposed to the web application through the *OpenCPU API* (Ooms 2014) and permits computation on corpus and lexicography resources; *rcqp* package (Desgraupes & Loiseau 2012) is used to connect to the *CQP* engine; A. Guerreau's scripts for lexical statistics (*Medialatinitas Github*) allow to find co-occurrences of the lemma in the corpus, while S. Evert's *wordspace* package (Evert 2014) is employed to calculate word similarities based on their distributional features.

2.2. Why to integrate: objectives?

medialatinitas.eu was created in order to:

1) improve access to distributed resources and facilitate dictionary writing process;

2) stimulate research on Medieval Latin vocabulary through linked resources and popularise innovative approach to the study of Latin text;

3) integrate community of experienced and early-stage researchers who want to apply computational methods in Latin philology, history and linguistics.

2.2.1. User's commodity

On the most intuitive level, handling scattered resources results in losing time and energy. This primarily stems from the very fact of the data being stored in different locations. Not only do the users have to consult multiple web pages, but they can never be sure the resource of their choice would be accessible, as its availability depends entirely on whether the distant service is actually running. Even if it is, each service or repository the user has to consult forces her to respectively adapt the search strategy, remind of the query syntax or verify integrity of the data. The latter, in particular, may be often difficult to asses, as many databases or text collections still lack appropriate documentation which would explain text or dictionary origin, its scope, data encoding scheme adopted and so on. In the worst scenario, scarcity of information would increase the disadvantages inherent in many non-research-driven web resources, that is lack of quality control, unclear or dubious choice principles, fragmentary and subjective character.

2.2.2. Answering old, asking new research questions

Yet, the user convenience, albeit important, is not the main objective of the *medialatinitas.eu* project. The principle that underpins the design of the present web application is to challenge the separation of knowledge components that should effectively cooperate in comprehension of the Medieval Latin text and culture. To achieve this goal and to compensate for the deficiencies each separate resource presents, *medialatinitas.eu* enables their concurrent, yet meaningful retrieval, and intertwine them in order to construct a coherent account of word's meaning potential, its grammatical and syntactical properties and cultural function. In the same time, *medialatinitas.eu* promotes alternative forms of access to linguistic data (charts, maps etc.) and their reuse in new research contexts.

At more specific level, *medialatinitas.eu* builds on linguistic content by addressing those issues which are either typical of lexicography description in general or affect the Medieval Latin dictionaries in particular, namely:

1) limited account of variation of the Latin vocabulary;

2) limited or inadequate frequency information which is based mainly on manual excerption of the linguistic evidence;

3) purely linguistic approach to sense definition.

Numerous benefits that come from closer integration of lexicographic and corpus resources need not to be enumerated here. Within the main interface of the

medialatinitas.eu corpus data are used to shed more light on the distributional properties of the Latin vocabulary. These are handled unsatisfactorily in the Medieval Latin dictionaries which did not adopt any coherent system of marking, for example, word frequency, except for such imprecise labels as 'more often' or 'very often'. This, in turn, makes distinguishing between widespread and limited phenomena often a challenging task, as the latter (*hapax legomena* included) are being traditionally given relatively more space than high-frequency lemmas. Moreover, existing evaluation of the frequency of word or grammatical/syntactical pattern is far from ideal, since it is based on evidence which was retrieved from the sources manually (Guerreau-Jalabert & Bon 2010; Bon, in print). The dictionaries also often fail to provide adequate account of the diachronic, diatopic and genological features of the word use. On the one hand, they would often overestimate stability of semantic or grammatical patterns through the ages, while neglecting their changing function and dynamic distribution across the text genres. On the other hand, the available dictionaries (part of them still in progress) cover neither all periods nor all geographical zones of Latin development. Targeted corpus query may compensate for their shortcomings in this regard.

Equally important are reasons for closer integration of encyclopaedic data. *medialatinitas.eu* draws on the research in modern linguistic theory which demonstrates that the distinction between linguistic competence and real-world knowledge is not as clearcut as the lexicographic practice shows (Geeraerts 2000). *medialatinitas.eu* searches, then, for a compromise between the rigour of purely linguistic definition and the fact that the users of historical dictionary usually need more information when trying to understand ancient text, as the amount of the shared cultural background is necessarily significantly limited. This is the more remarkable, as Medieval Latin was for centuries the language of scientific, theological and philosophical writing, so exhaustive dictionaries (as majority of currently published are) inevitably have to deal with this terminological richness[2]. Although medieval terminology calls for different sense defining strategy than one applied in general lexicography, one often comes across definitions that, due to their purely linguistic character, are virtually void of any explicative potential. Meaningful reuse of encyclopedic resources in the *medialatinitas.eu* application will help to tackle such specific cases and enrich dictionary content in general.

Finally, closer integration of encyclopaedic and lexicographic data is desired for practical reasons. A good example in that respect might be proper names which are traditionally excluded from Medieval Latin general dictionaries. Yet, the correct decoding of place or personal names is crucial for understanding ancient text and constructing its referential layer. As a result, the readers of a medieval author will often find themselves consulting dictionaries and encyclopaedias at the same time. Apart from the user convenience, however, including proper names will be of

2It makes some researchers claim that the Medieval Latin language was practically a special language (Bon 2013).

benefit, for instance, when describing common nouns if the latter are motivated by the former or vice versa (e.g. *aqua* 'water' is a component of many place names) etc.

2.2.3. Building community of users and developers

Finally, the present work aims at integrating community of the developers and researchers. Although there now exists an active community of digital medievalists and the number of researchers who apply computational methods in their work on medieval texts has been steadily growing, until now no large-scale effort has been undertaken in order to integrate distributed data or to help developers to embed their code snippets into a larger application. The same is true about pedagogical resources: despite numerous individual initiatives that have been taken (e.g. on-line bibliographies etc.), the researchers willing to exploit automatic methods in their work cannot refer to any set of guidelines which would be appropriate for Latin text processing and query. This is why *medialatinitas.eu* will enable users to contribute their widgets[3] as R and JavaScript code snippets responsible for single, yet self-contained functionality. Finally, the knowledge base that will constitute an important part of the *medialatinitas.eu*[4] will provide users, on the one hand, with a curated collection of guidelines, showcases and links, and, on the other hand, with a complete description of digital medievalists' workflow - from the OCR to the corpus query.

2.3. How to integrate: application design and architecture

In the current development stage, *medialatinitas.eu* sticks with an integration model that could be characterised as 'shallow'. The word is, however, used in the pregnant sense, as it is meant to describe implementation which is shallow, lightweight and agile at the same time.

The present integration model is called *shallow*, firstly because the data are not provided any explicit links and the integration takes place principally in the application user interface (UI). As for now, virtually no effort has been put into harmonising different classes of the resources, also same-class data are stored or dynamically queried „as they are". Dictionaries, corpora and encyclopedias do not refer to any common system of identifiers, therefore, for example, there is no formal connection established between the dictionary headword AQUA 'water', the lemma AQUA in the annotated corpus and the *Latin Wikipedia* article for AQUA. As was already said, the same applies for same-class data, so, for instance, there is no inherent mapping between the entry AQUA in the DuCange's *Glossarium mediae et infimae latinitatis* and its equivalent in the *LMILP*; similarly, there exists no explicit link between two identical lemma labels in separate corpora, if they have been annotated with different lemma sets. *Ad-hoc* equivalence between two

3See below.
4The knowledge base is beyond the scope of the present paper.

dictionary headwords or lemmatised word forms is established if they have an identical orthographic form and share a PoS label. Other resources are currently retrieved based on a simple full-text query.

Secondly, *medialatinitas.eu* is designed as a three-level deep application (Figure 3) offering for each lemma the user may look up: 1) a general overlook; 2) an extended view; 3) an advanced view.

Figure 3: Three levels of the *medialatinitas.eu* application: 1) general view; 2) extended view; 3) native application (here, *CQPweb*).

1) When visiting the main page, the user initially comes across a simple search form. Once the query phrase is specified (currently only lemma search is supported), it is next despatched to locally and remotely running services and APIs. The returned results are processed and, subsequently, displayed on the same page.[5] Its layout is built around a grid system and consists of a series of separate widgets, each responsible for displaying some portion of information about the word in question. As a whole, the widgets contribute to a general, yet varied outlook of the word meaning, its linguistic properties and distribution. The widgets that have been implemented so far present, for instance: 1) short excerpts from

5Single Page Application (SPA) model of Web application design is here adopted.

definitions of the Classical and Medieval Latin dictionaries; 2) short extracts from corpora concordances; 3) selected morpho-syntactic properties (inflectional type, gender, tense or case endings *etc.*) of the word (retrieved from the electronic dictionaries); 4) distribution of word forms in the corpora; 5) diachronic and genological distribution of the lemma; 6) co-occurrent terms in selected corpora; 7) similar words in selected corpora; 8) translations and similar terms (retrieved from the *BabelNet*); 9) links to the *Latin Wikipedia* pages whose text contains word in question; 10) list of quotations which contain the searched lemma (retrieved from the *Latin Wiktionary*); 11) list of titles of literary works which contain the lemma (retrieved from the *Internet Archive*); 12) list of images (Figure 4) whose description contains the lemma (retrieved from the *Europeana*); 13) map of the place names (Figure 5) that contain the lemma (retrieved from the *Pelagios Project*, *Getty Thesaurus of Geographic Names* and *GeoNames*).

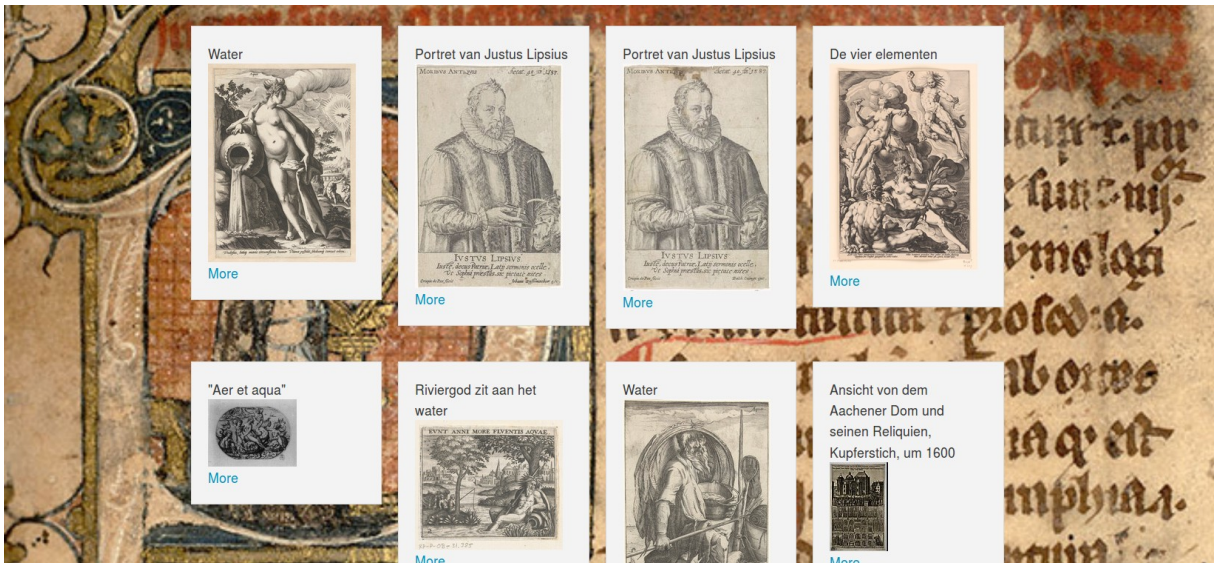

Figure 4: Media widget: images whose description matches the string *aqua* 'water' (fetched from the *Europeana*).

medialatinitas.eu employs various forms of data display, widgets are, thus, implemented as tables (1-3, 8) or lists (9-12), but also as charts, visualisations (4-7) and maps (13).

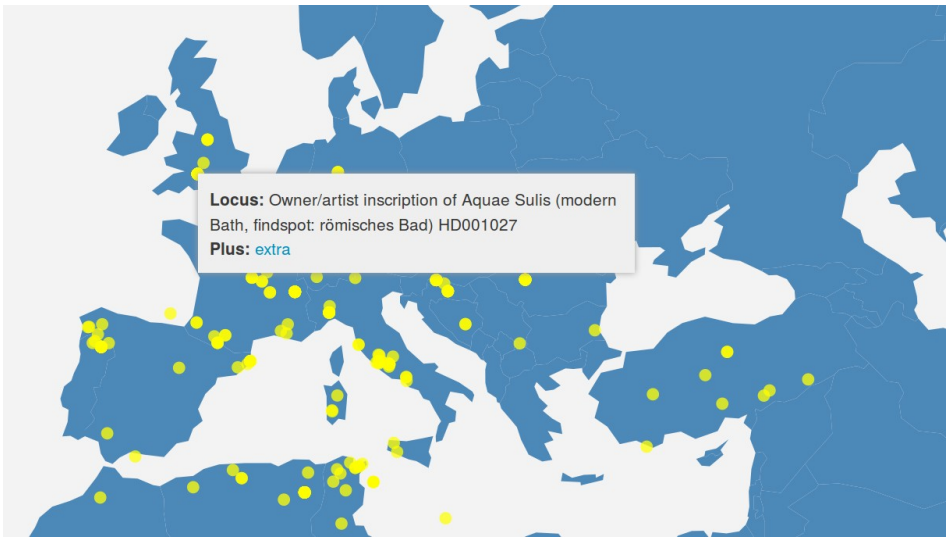

Figure 5: Map widget: yellow points represent ancient place names composed of the lemma *aqua* 'water' (geographical coordinates and labels are fetched from the *Pelagios Project API* and visualised with the *d3.js* library)

II) the extended view, which is beyond the scope of the present paper, is accessible upon clicking on any of the main page widgets and offers more detailed and focused perspective on the selected properties of the lemma. For the moment, only *shiny*-based (Chang et al. 2015) dashboard for lexical statistics has been developed.

III) the native application (*CQPweb*, *eXist-db*, *R shiny* interface etc.) is accessible from the extended view and constitutes the deepest layer of the *medialatinitas.eu* web interface.

The present application can be considered as *lightweight*, as it has no pretensions to become a complete virtual research environment. It is conceived as a modular platform that would allow to plug in rapidly new widgets and test dynamically alternative modes of linguistic data representation. Because at last it always refers the users to a native application, there is no ambition to replace existing software, as it is believed that such mature tools as, for instance, *CQPweb* offer exceptional set of features that one can effectively build on. As a result, *medialatinitas.eu* is *agile*[6], since it is open to further expansion and should change according to the research interest and needs of its users and developers.

3. Discussion

3.1. *medialatinitas.eu* as a mashup application

In its design principle, *medialatinitas.eu* is closest to a mashup which is defined as „a composite application developed starting from reusable data, application logic, and/or user interfaces typically, but not mandatorily, sourced from the Web" (Daniel & Matera 2014, 3). It is 'composite', as it integrates data from more than one Web services, each of which is a full-blown web application or a SPARQL

6The use of this term is distantly inspired by the notion of *agile software development*.

endpoint. Following Daniel and Matera's classification, *medialatinitas.eu* may be further described:

- regarding its „composition", as a hybrid mashup, for the integration takes place both in the application logic and in the UI layer;

- regarding its „domain" or purpose, as a scientific, discovery-driven mashup;

- regarding „environment" or „deployment context", as a Web mashup in which logic layer is distributed over client and server: whenever small data portion is involved, the client application written in AngularJS is responsible for processing Ajax calls, computing on their results and presenting the results; however, once larger datasets come into play, especially when user changes from the general view to the more specific one or when heavy calculation is to be applied, the burden of processing shifts towards the server and the client takes only care of visualising the returned data.

User's lemma query is passed to a mediator which subsequently transfers it to a series of wrappers. These, in turn, execute API calls and return back the results. The mediator, then, tackles the syntactic heterogenity of the data, while wrappers deal with idiosyncrasies of each source, thus resolving schematic heterogenity. Problem of semantic heterogenity of the data remains, as was already mentioned, unresolved and needs to be addressed in the nearest future by compiling a canonical list of lemmas that could be used to harmonise headwords of dictionaries, corpus annotations and encyclopedic entities.

3.2. Meaningfulness, narrative and reproducible research

Rather than only assemble pieces of information in one place, *medialatinitas.eu* aims at providing whenever possible a relatively exhaustive and coherent narrative of each lemma. As for the exhaustiveness, the variety of the resources employed assures that no crucial level of word description is omitted. Dictionaries, apart from obvious semantic, provide also morphological, orthographical, syntactical and pragmatical information. Corpora contribute to the description of frequency, collocational features and computed meaning of the word. They are also a valuable source of knowledge about diachronic evolution of the lemma. Finally, the cultural component is covered by use of paremiological resources, iconographic evidence (which helps to trace down allegorical sense), thesauri (for example, plant names) *etc.*

The idea that *medialatinitas.eu* should provide its users with a coherent and meaningful narrative has at least three sources. Firstly, it is a reaction to the growing popularity of automatically compiled on-line content aggregators in which the very fact of juxtaposing multiple resources seems often to suffice as their *raison d'être*. Such seemingly objective form of data presentation, at the same time,

obscures the fact that the composition itself is already an interpretation. Secondly, the presence of contextualising, explicative or interpretive commentary seems to be what may distinguish human-oriented research applications from the popular, yet mainly machine-oriented resources, such as *WordNet* or *BabelNet*. Thirdly, *medialatinitas.eu* is also an exercise in new form of lexicographic discourse in the era of the linked linguistic data.

At the most basic level the narrative „glue" is generated in form of short introductory phrases which precede each widget or widget group. Being functionally equivalent to the headers, they do not add any substantial information, but, first, enable users to get instant insight into what linguistic or cultural phenomenon is represented in a specific section of the page and, secondly, make possible reading the whole page as a continuous text.

Apart from that, the narrative is built across the page by means of three other devices:

1) graphical and textual hints;

2) explicative and interpretative passages;

3) dynamically generated reports.

Graphical and narrative hints that the users find all over the interface indicate quality, scope and completeness of the presented data. Since *medialatinitas.eu* is to be a research tool, the users need to be able to assess, first of all, whether they may safely draw conclusions from the gathered resources, and, secondly, whether insights offered in visuals, such as maps or charts, are of more than decorative value. To this goal, graphical signs and corresponding labels have been employed throughout the page which signal:

- whether a widget was built on a resource which is of high, low or unknown quality;

- which chronological and geographical dimension a specific resource represents and

- whether it covers some phenomenon fully or only partially.

In a practical case of an excerpt from the *LMILP*, the quality, scope and coverage would be set *resp.* as „high (academic)", „10-15th c., Poland", and „full", whereas in the case of an OCR-ised text they would be specified as „low (OCR)", „6-12th c., Europe" and „partial".

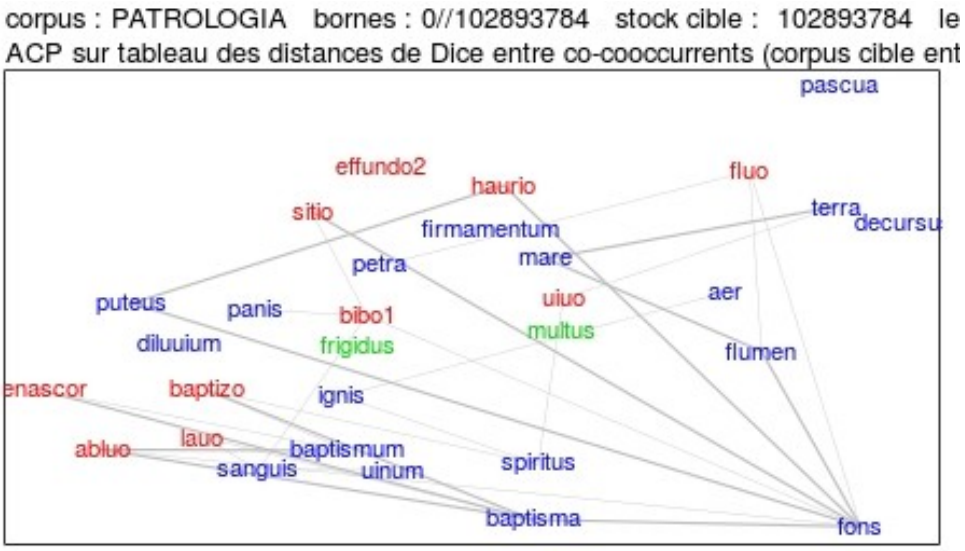

Figure 6: PCA chart: computed co-occurrences of the lemma *aqua* 'water' in the *Patrologia Latina* corpus (generated with A. Guerreau's R script).

In the *medialatinitas.eu* visualisation widgets are accompanied by a short passage whose role is to explain what procedures have been applied to yield the results and to help at their interpretation. There are at least two reasons for providing such explanation. First of them is that *medialatinitas.eu* sticks with the reproducible research paradigm. At any time, the user may learn not only how specific visualisation was generated, but also explore its theoretical background. Secondly, some less standard forms of data presentation cannot simply do without commentary text, if they are to be more than a decorative device. Whereas a barplot illustrating diachronic distribution of a specific word is relatively self-explanatory, the same cannot be said about the boxplots, PCA charts (Figure 6) or co-occurrence barplots (Figure 7) which should be accompanied by a supplementary text if they are not to overwhelm a less advanced user.

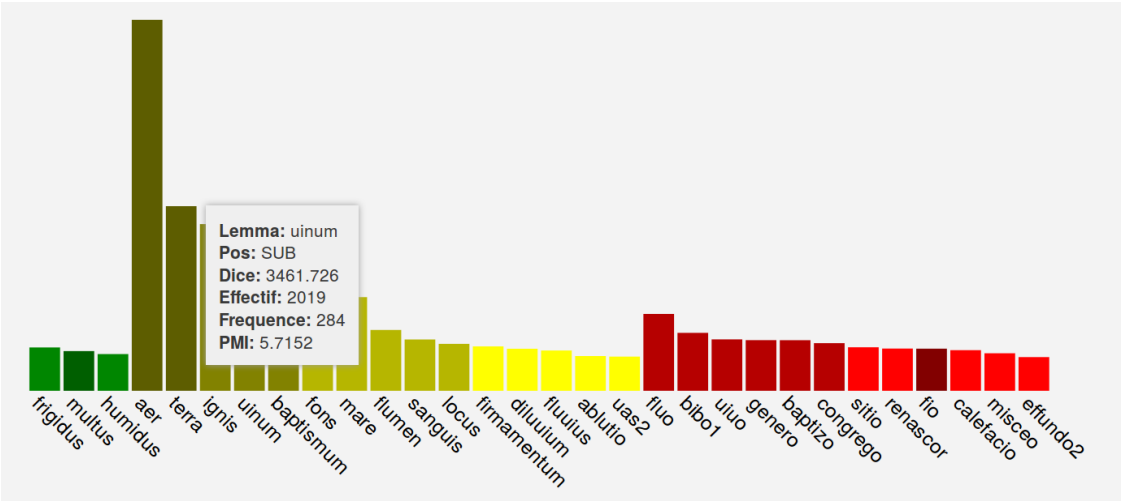

Figure 7: Barplot representing computed co-occurrences of the lemma *aqua* 'water' in the *Patrologia Latina* corpus (data fetched from an R session exposed with OpenCPU API; the chart generated with the help of the *d3.js* library).

Hints are, therefore, provided as to how one can interpret the geometric properties of the chart, such as distance between the points, the width of the boxplot, and so on. In the case of the co-occurrence barplot, for instance, apart from the information provided in the legend, one may learn that the selected coefficient promotes some type of collocates or that the intensity of the color of the bar corresponds to the absolute frequency of the co-occurrent in the specific corpus.

Finally, the paradigm of reproducible research is further promoted by enabling users to download complete reports from their queries. Since the reports are available not from the main, but from extended view page built with *shiny R* and *OpenCPU*, they will not be explained here in more detail.

4. Conclusion

4.1. Previous research

The integrative, holistic approach to the word meaning has been a central idea of the philology since its Greek origins, but is also accentuated in modern, cognitive lexical semantics.[7] The architecture of presented application, which is a hybrid tool (Granger 2012), benefits from research on mash-ups and content aggregation (Daniel & Matera 2014). *medialatinitas.eu* makes heavy use of visualisation techniques and other alternative ways of lexicographic data representation and in that respect it builds on recent research into linguistic data visualisation (Theron & Fontanillo 2013). The notion of reproducible research in scientific computing has recently gained interest, as easy to use R packages such as *knitr* (Xie 2014) were made available. Although, at the current stage, *medialatinitas.eu* adopts a lightweight, UI-based model of data integration, further work will certainly focus on closer data integration following the LLOD model (Chiarcos et al. 2013). Already in its present form, though, the application exploits existing Semantic Web resources[8], such as *BabelNet*, *Europeana*, *Getty Thesaurus of Geographical Names*, *Pelagios Project* etc. Integration of the lexicographic resources is promoted and stimulated to an unprecedented extent within the *European Network of e-Lexicography* (*ENeL*) of which members are the authors of the present paper.

Interest in dictionary applications that follow the aggregator or mash-up model seems to be rapidly growing in recent years, with such eminent examples as *Dictionary.com*, *FreeDictionary*, or *Wordnik*.[9] Regardless of the reasons for it, the aforementioned resources usually offer aggregation of the dictionary content within

7Geeraerts (1988; 2009) is one of few researchers to notice the link between the historical-philological and cognitive semantics.

8The extensive list of dictionary APIs can be found on the*ProgrammableWeb* website. Accessed at: http://www.programmableweb.com/category/all/apis?keyword=dictionary. (23 May 2015)

9According to the *Alexa* website ranking, among 20 most viewed dictionary pages there are at least three resources of this kind: *The FreeDictionary* (the 380th most popular page on the web and the second most popular dictionary after *WordReference*), *SpanishDict* (1827th) and *Your Dictionary* (2116th).

a user-friendly interface equipped with efficient query engine. Yet, in the majority of cases they reuse popular general dictionaries, do not offer any further commentary concerning fetched data and in particular they do not inform about the credibility of the resources. This makes them hardly usable as research tools. The situation is slightly different when one takes into account such aggregators as *Dictionnaire vivant de la langue française*. One will find here juxtaposed the excerpts from renown lexicographic works (e.g. *TLFi*), but also selection of corpus and web quotations, as well as charts illustrating changing frequency of the word. The *DVLF* seems to adopt the same design as that of *Logeion* which, apart from aggregating Latin dictionaries, presents for each lemma additional information based on the *Perseus Digital Library*: list of authors who frequently use the word and a small selection of co-occurrent terms. *medialatinitas.eu* differs from the websites mentioned above not only in the general architecture or scope of the integration, but also in the resources employed, use of encyclopaedic data, implementation of complex statistics, visualisation techniques etc. The same properties distinguish *medialatinitas.eu* also from more general oriented text analysis framework such as the *Perseus Digital Library* which collects a large number of lemmatised Latin and Greek texts of Classical Antiquity and Renaissance. Apart from the already mentioned differences, *medialatinitas.eu* is principally lemma-, not text-oriented, therefore it is expected to be used as a tool for exhaustive analysis of the vocabulary and not as a reading environment. Moreover, *medialatinitas.eu* employs graphical hints, rich visualisations and mapping, exploits modern academic works rather than older dictionaries or text editions, goes beyond in-house resources and uses transparent cooccurrence and frequency measures. Otherwise than the *Perseus*, it also makes clear distinction between text collection and linguistic corpus and contains a great deal of medieval texts. Contrary to the *Perseus* which does not seem to evolve much for the last few years, *medialatinitas.eu* is conceived as a modular, open to extension, lightweight application.

4.2. Further development

The future development of the *medialatinitas.eu* will focus on four main objectives. First, more appropriate model of linguistic data integration needs to be adopted in order to better deal with the diachronic evolution of Latin vocabulary and with conflicting annotations of linguistic resources. Apart from faster and more direct search, closer integration should also lead to more sophisticated processing of the user's input. Currently, the search is limited to lemmas only and as such it requires from the user rather good knowledge of the Latin language. Secondly, more data should be hosted locally which should help to lower the query and page display times. It is also desirable, because the external APIs (the *BabelNet HTTP API* is one example) often limit number of queries that can be sent from a single IP address. Thirdly, new widgets should be added and the existing ones need to be constantly improved. The system of graphical hints should be refined and more

techniques of data representation and computation should be suggested. Finally, community of users and content providers needs to be expanded.

5. Acknowledgements

Work on the present paper has been funded by:

1) the „Young Reasearcher Grant" of the Polish Ministry of Science and Higher Education attributed to Krzysztof Nowak by the Institute of Polish Language (Polish Academy of Sciences);

2) the "Soutien à la mobilité internationale" grant attributed to Bruno Bon by the Institut des Sciences Humaines et Sociales (Centre National de la Recherche Scientifique).

The paper has greatly benefited from the discussions and workshops of the *ENeL* COST Action in which both authors have opportunity to participate.

The authors would also like to thank Renaud Alexandre (IRHT CNRS) for his remarks.

6. References

Alexa. Accessed at: http://www.alexa.com/. (23 May 2015)

BabelNet. Accessed at: http://babelnet.org/. (23 May 2015)

Blatt, F. & Lefèvre, Y. & Monfrin, J. & Dolbeau, F. & Guerreau-Jalabert, A. (eds.). (1957-2011). *Novum Glossarium Mediae Latinitatis.* Copenhague/Bruxelles/Genève. Available at: http://www.glossaria.eu/ngml.

Bon, B. (2013). Le vocabulaire technique en latin médiéval, entre mythe et réalité. In H. Leithe-Jasper & M.-L. Weber (eds.) *Fachsprache(n) im mittelalterlichen Latein / Technical Language(s) in the Latin Middle Ages / Langage(s) technique(s) au moyen âge latin, Tagungsakten der fünften internationalen mittellateinischen Lexikographentagung (München, 12.-15. September 2012). Archivum Latinitatis Medii Aevi*, 71, pp. 355-375.

Bon, B. (in print). Histoire et perspectives du 'Novum Glossarium Mediae Latinitatis'. *Proceedings of the 7th International Conference on Historical Lexicography and Lexicology (ICHLL 2014)*. Bern/New York: Peter Lang.

Bon, B. & Nowak, K. (2013). WikiLexicographica: Linking Medieval Latin Dictionaries with Semantic MediaWiki. In I. Kosem & J. Kallas & P. Gantar & S. Krek & M. Langements & M. Tuulik (eds.) *Electronic Lexicography in the 21st century, Thinking outside the paper: Proceedings of the eLex 2013 Conference*. Tallinn-Ljubljana: Trojina, Institute for Applied Slovene Studies; Eesti Keele Instituut, pp. 407-420. Available at: http://eki.ee/elex2013/proceedings.

Chang, W. & Cheng, J. & Allaire, JJ. & Xie, Y. & McPherson, J. (2015). *shiny: Web Application Framework for R.* Available at: http://CRAN.R-project.org/package=shiny.

Chiarcos, C. & McCrae, J. & Cimiano, Ph. & Fellbaum, Ch. (2013). Towards Open Data for Linguistics: Linguistic Linked Data. In A. Oltramari & P. Vossen & L. Qin &. E. Hovy (eds.) *Theory and Applications of Natural Language Processing.* Berlin/Heidelberg: Springer, pp. 7–25.

Corpus Thomisticum. Accessed at: http://www.corpusthomisticum.org/it/index.age. (23 May 2015)

Croatiae Auctores Latini. Accessed at: http://www.ffzg.unizg.hr/klafil/croala/. (23 May 2015)

d3.js. Accessed at: http://d3js.org/. (23 May 2015)

Daniel, F. & Matera, M. (2014). *Mashups: concepts, models and architectures*, New York: Springer.

Desgraupes, B. & Loiseau, S. (2012). *rcqp: Interface to the Corpus Query Protocol.* Available at: http://CRAN.R-project.org/package=rcqp.

Dictionary.com. Accessed at: http://dictionary.reference.com/. (23 May 2015)

ENeL. European Network of e-Lexicography. Accessed at: http://www.elexicography.eu/. (23 May 2015)

Europeana. Accessed at: http://www.europeana.eu/portal/. (23 May 2015)

Evert, S. (2014). Distributional Semantics in R with the wordspace Package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations.* Dublin: Dublin City University/Association for Computational Linguistics, pp. 110–114. Available at: http://anthology.aclweb.org/C/C14/C14-2024.pdf.

eXist-db. Accessed at: http://www.glossaria.eu/treetagger/. (23 May 2015)

Fontes. Corpus of Polish Medieval Latin. Accessed at: http://scriptores.pl/fontes. (23 May 2015)

Geeraerts, D. (1988). Cognitive Grammar and the History of Lexical Semantics. In B. Rudzka-Ostyn (ed.) *Topics in Cognitive Linguistics.* Amsterdam: John Benjamins Publishing Company, pp. 647-677.

Geeraerts, D. (2009). *Theories of Lexical Semantics.* Oxford: Oxford University Press.

Geonames. Accessed at: http://www.geonames.org/. (23 May 2015)

Getty Thesaurus of Geographic Names. Accessed at: http://www.getty.edu/research/tools/vocabularies/tgn/. (23 May 2015)

Glossarium mediae et infimae latinitatis. Accessed at: http://ducange.enc.sorbonne.fr/. (23 May 2015)

Granger, S. (2012). Introduction: Electronic lexicography - from challenge to opportunity. In S. Granger & M. Paquot (eds.). *Electronic lexicography.* Oxford: Oxford University Press, pp. 1–11.

Guerreau-Jalabert, A. & Bon, B. (2010). Le trésor au Moyen âge: étude lexicale. In L. Burkart & al. (eds.) *Le trésor au Moyen âge.* Firenze: Sismel, pp. 11-31.

Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17 (3), pp. 380–409.

Internet Archive. Accessed at: https://archive.org/. (23 May 2015)

Krötzsch, M. & Vrandečić, D. & Völkel, M. & Haller, H. & Studer, R. (2007). Semantic Wikipedia. *Journal of Web Semantics* 5 (4), pp. 251–61.

Latin Wiktionary. Accessed at: http://la.wiktionary.org/wiki/Pagina_prima. (23 May 2015)

Le Dictionnaire vivant de la langue française. Accessed at: http://dvlf.uchicago.edu/. (23 May 2015)

LMILP: eLexicon Mediae et Infimae Latinitatis Polonorum. Accessed at: http://scriptores.pl/elexicon. (23 May 2015)

Logeion. Accessed at: http://logeion.uchicago.edu/. (23 May 2015)

Medialatinitas Github. Accessed at: https://github.com/medialatinitas/. (23 May 2015)

Nowak, K. (2014). The eLexicon Mediae et Infimae Latinitatis Polonorum, Electronic Dictionary of Polish Medieval Latin. In A. Abel & C. Vettori & N. Ralli (eds.) *The User in Focus: Proceedings of the XVI EuraLex International Congress*. Bolzano/Bozen, pp. 793-806. Available at: http://euralex2014.eurac.edu.

Omnia Project Treetagger. Accessed at: http://www.glossaria.eu/treetagger/. (23 May 2015)

Ooms, J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. *ArXiv e-prints*. Available at: http://arxiv.org/pdf/1406.4806v1.pdf.

Open Library. Accessed at: https://openlibrary.org/. (23 May 2015)

Orbis Latinus. Accessed at: http://olo.rigeo.net/. (23 May 2015)

Pelagios Project. Accessed at: http://pelagios.dme.ait.ac.at/api. (23 May 2015)

Perseus Digital Library. Accessed at: http://www.perseus.tufts.edu/hopper/. (23 May 2015)

R Core Team (2015). *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester. Available at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Semantic Mediawiki. Accessed at: https://semantic-mediawiki.org/. (23 May 2015)

SpanishDict. Accessed at: http://www.spanishdict.com/. (23 May 2015)

The Free Dictionary. Accessed at: http://www.thefreedictionary.com/. (23 May 2015)

Theron, R. & Fontanillo, L. (2013). Diachronic-Information Visualization in Historical Dictionaries. *Information Visualization* 14 (2), pp. 111–36.

TLFi: Le Trésor de la Langue Française Informatisé. Accessed at: http://atilf.atilf.fr/tlf.htm. (23 May 2015)

VIAF. Accessed at: http://viaf.org/. (23 May 2015)

Wordnik. Accessed at: https://wordnik.com/. (23 May 2015)

Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch, & R. D. Peng (eds.) *Implementing reproducible research*. Boca Raton: Chapman and Hall/CRC, pp. 3-32.

YourDictionary. Accessed at: http://www.yourdictionary.com. (23 May 2015)