Progressive Advancement of Knowledge Resources and Mining: Integrating Content Factor and Comparative Analysis Methods for Dynamical Classification and Concordances

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—The research presented in this paper concentrates on the results from creating new advanced methodologies used for enhancing knowledge resources and knowledge mining. Creating and developing advanced knowledge resources and features for mining over long-time periods are challenging tasks, which require the continuous development of advanced complex means. Enhancements have to include semi-manual and automatable implementations from advanced methodologies in order to access the knowledge- and context-related characteristics and context values. This research builds on the practice of creation and development of multi-disciplinary knowledge resources for decades and creating and applying mining and discovery methods. The results on creating means for a comparative analysis of data entities from knowledge resources are used for data analysis and enhancement of complex and increasing multi-disciplinary knowledge resources. Comparing data entities is a most ambitious task for increasingly complex data objects, integrated resources, and relations – from the Knowledge Resources, as well as from the computational perspective. The implementation utilises complementary components, which enable to structure and describe complex knowledge and support an advanced analysis. Based on the methodological fundament, the paper presents practical results, delivers and discusses instructive examples from an implementation and case study. For practical reasons with the comparative analysis, the knowledge resources explicitly utilise references to publicly available data resources. The goal of this research is the enhancement of knowledge resources and mining by modular implementations of advanced methodologies, especially method integration for comparative analysis and knowledge mining with information systems and long-term multi-disciplinary knowledge resources.**

***Keywords*–Enhancement of Knowledge Resources; Comparative Analysis; Content Factor; Universal Decimal Classification; Advanced Data-centric Computing.**

I. Introduction

It is a truth universally acknowledged, that knowledge and knowledge-related data are core values of human activities. Knowledge and long-term knowledge resources are treasure troves of documentation and sources of major insights. The enhancement of knowledge resources and documentation also contributes to new discovery and insights. Therefore, new advanced methods need to be created contributing to the development and enhancement of knowledge resources. The fundaments of the method providing a Comparative Analysis of data entities were presented at the INFOCOMP 2017 conference in Venice, Italy [1]. This paper presents results of extended research, which showed to be relevant for the enhancement of knowledge resources and mining and discovery processes. The enhancements include further aspects for integration of resources, the integration of conceptual knowledge, the architecture of the implementation, including the integration of different categories of resources, and computational aspects.

Advanced methods of knowledge mining with information systems and knowledge resources are becoming increasingly important. With that, improving knowledge mining and at the same time integrating larger amounts of data increases the challenges. The core of challenges is the data analysis. Within data analysis, comparing "data" is a central task. Comparing data entities is an even more ambitious task when data objects and relations are becoming more and more complex.

The term data entity in context with knowledge resources refers to any data representing objects of any kind like digital or realia objects, including references, e.g., to objects or conceptual knowledge. Within this research, special application components were created and implemented in order to provide modular means to be integrated for a comparative analysis, e.g., knowledge resources referring to structured and unstructured data, conceptual data, especially knowledge classification, and methods specialised on the before mentioned means, e.g., the Content Factor method (CONTFACT) [2]. It is not the task of this practical implementation to re-iterate the basics of the instruments used. The basics of the method are explained in theory and practice in the original publication as well as theory, definitions, and context are explained and defined in referred publications.

The multi-disciplinary knowledge resources and the application of the Content Factor method have enabled new flexible workflows and the creation of new complementary means for both the enhancement of multi-disciplinary knowledge resources and for data-centric knowledge mining and discovery processes [3]. Some of the most widely required means with data entities of knowledge resources are components for a comparative analysis. Comparative Analysis (CA) is defined as an item-by-item comparison of two or more comparable entities. The goal of this research is the enhancement of knowledge resources and mining by appropriate integration of methods.

Resulting from this research, the enhancements are based on the new and original integration of Comparative Analysis and the Content Factor methods, an integration which is implemented and approved with the research projects in various practical context.

This paper is organised as follows. Section II summarises the state-of-the-art, motivation, and frame of reference to the ground of comparison. Section III introduces to knowledge and integration of different resources, as implemented for this case study. Section IV presents the integration of conceptual knowledge and Sections V and VI show the knowledge-centric integrative architecture for the computation and analysis of results based on the selected resources and the implementation of the workflow. Section VII discusses the main results and evaluates them in context of the application. Section VIII summarises the results and lessons learned, conclusions, and future work.

II. State-of-the-art, motivation, and frame

The elementary way of knowledge mining, practised by the vast majority of approaches and services ignores content quality, document types, and cognitive knowledge. That means, content is handled independently from the creation process and expertise, content from databases, Web pages, and scanned books are not differentiated, and classification of content is disregarded.

CA modules can be used for arbitrary purposes with knowledge mining workflows, e.g., for selecting complementary resources as well as selecting objects supporting decision making processes. The method is used with knowledge mining workflows, integrating dedicated knowledge resources and publicly available content, e.g., text documents and books, because of their complementary nature regarding content, structure, and quality.

The complexity requires to start with a comprehensive high level view and context for the target of this research. The following sections describe the motivation and the base of the conducted CA.

A. *Frame of reference*

The significance of integrating different data entities results from the context, in which they are placed. This research presents a method of comparing different data entities as referred from objects in advanced knowledge resources.

Objects with higher quality are mostly more complex. Advanced knowledge mining and decision making requires more than one method or algorithm for analysis of available objects and their references, data entities, and attributes. A major challenge is the difference of entities, e.g., regarding entity type, original purpose of the entity, and source but also content and structure.

Different types of entities cannot be ignored from advanced workflows because they contain unique knowledge and information. In most cases, the knowledge and information can even only be provided by different entities and referred sources. Methods should be provided, which are beneficial to be integrated in advanced workflows, especially for analysis, quantisation, and qualification of different entities. The deployed means should allow long-term data-centric applications and intrinsically foster the seamless integration with existing workflows. In addition, the methodologies, methods, and architecture of integration should allow the implementation of modular and least invasive components.

B. *Grounds for comparison*

Besides the complexity, a combination of data entities from different sources and different types was choosen for the following reasons. The rationale behind the choice for knowledge resources and entities from referred objects results from complementary content and context. There is an arbitrary high quality of multi-disciplinary content in the knowledge resources, which are in continuous development [4]. In addition, the knowledge resources can provide an extremely high knowledge and information density. The Gutenberg resources [5] can provide a large number of fully publicly available standard text documents and elaborations for a wide multi-disciplinary context. Both types of resources contain essential amounts of textual content and are continuously extended and improved. The relationship between different entities is the addressed knowledge content with its unique nature. The thesis is, that different entities should neither be left out from advanced workflows nor should their content, the unique knowledge and information, be ignored.

C. *Organisational scheme*

The targeted lens comparison discusses the most important aspects text-by-text, focussing on advanced knowledge resources and referred resources.

For complex lens comparisons of that kind we require to have an extended focus on integration of resources, including aspects of automatically created and integrated objects. The integration further has to resolve and manage conceptual knowledge, especially for important sources of conceptual knowledge. Creating concordances is a common means used with classifications. Concordances are used for providing references between classifications, they are not used for comparing or analysing content or data. With the overall goal of enhancing knowledge resources and mining, supporting and generating concordances are major means for achieving the integration of conceptual knowledge. Special care is taken for aspects of computation and analysis, due to the large complexity and data and the requirements for compute intensive advanced methodologies and algorithms.

III. Knowledge and the integration of resources

The following sections describe how an integration was achieved and which results were gained with the analysis.

A. *Knowledge*

One of the most important sources of understanding knowledge is Aristoteles' treatise of the Nicomachean Ethics [6], which is a classic basis of its essence [7], significance, and terms [8].

If one re-visits a place it will, to some extent, be a different place [9]. There are long-term and short term changes. Everything in the world is connected. This is also true for knowledge resources, embedded in existing context. Knowledge resources contain multi-disciplinary knowledge objects, which can be used in arbitrary ways for providing knowledge, e.g., factual, conceptual, procedural, and metacognitive knowledge [10].

The objects [11] can contain any content and context as well as references, e.g., translations, transliterations, synonyms, associations [12], references [13], conceptual knowledge (e.g., UDC), concordances [14], links, references (see, "s."), optional references (see also, "s. also"), comparable references (compare, "comp."), keywords, and Content Factors (of elements) [15]. The objects can be based on records, e.g., characters, words, lines, and complex records. In practical application scenarios [16], any content and context can be used for analysis and evaluation of an object.

B. Data entities

Data entities can be created from many resources. With this research, knowledge objects and data entities were automatically created from 'Gutenberg documents'. At the time of the case study (January 2017) Project Gutenberg [5] offered 53,855 free ebooks for download. At the time of preparation of this article (February 2018) Project Gutenberg offered 56,432 free ebooks for download. The document files include the text in a version of the respective edition, which can be a revised edition or translation. The text editions are linked as different document files, e.g., plain text files, which can be converted into data entities and integrated with different data entities. Regarding conceptual knowledge, the Gutenberg documents use a flat implementation of the Library of Congress (LoC) classification outline [17]. The ebook links contain some relevant information, too, e.g., the bibliographic record, EBook-No. 25062, a link the LoC Class entries, and the release date of the edition. The original publication date of the source text is contained in the document files. Publications like books and articles are not "unstructured". They differ in structure from knowledge resources' content but they can be seen as structured entities, too. Data entities from knowledge resources' collections and containers [2] are used with many knowledge mining applications [4].

With the created modules the data entities from the Gutenberg resources can be handled in the same way as knowledge resources' objects, e.g., of different origin. The following case study starts with a knowledge mining request for "Vesuvius" in the context of "volcanology". The primary Gutenberg result matrix contains a number of documents [18]-[24] provided with the precomputation. The essential steps and data of the examples should be comprehensible as these choosen resources are publicly available.

C. Object and data entity integration: Four cases

The following passages introduce four different types of objects, which were considered in order to integrate the available knowledge. The four types of objects are originating from three major groups of resources, namely publicly provided book and document object, collection objects, and container objects. For the first group, it is shown that the respective objects can have significant different characteristics, which is shown comparing two samples.

Objects and data entities can be integrated with knowledge resources in arbitrary ways, e.g., as a referred object or by creating an instance of an object. Here, for the goal of this research, the programming and programming languages of the modules are not relevant for the demonstration. Significant is the integration, which allows an analysis and evaluation, e.g., with knowledge resources' objects. The following excerpt (Figure 1) shows a knowledge resources' object automatically created from an entity of Gutenberg document 33483 [19] with LoC classification [25].

```
33483-0.txt [Document]:
            ...
            THE
            ERUPTION OF VESUVIUS
            IN 1872, ...
            BY
            PROFESSOR LUIGI PALMIERI,
            _Of the University of Naples; Director of the Vesuvian Observatory._
            ...
            WITH NOTES, AND AN
            _INTRODUCTORY SKETCH OF THE PRESENT STATE OF KNOWLEDGE_
            OF
            TERRESTRIAL VULCANICITY,
            _The Cosmical Nature and Relations of
            Volcanoes and Earthquakes._
            ...
            BY
            ROBERT MALLET,
            _Mem. Inst. C.E., F.R.S., F.G.S., M.R.I.A., &c., &c._
            ...
            WITH ILLUSTRATIONS. ...
            LONDON: ...
            _ASHER & CO._,
            13, BEDFORD STREET, COVENT GARDEN, W.C. ...
            1873.  ...
            W. S. Johnson, Nassau Steam Press, 60, St. Martin's Lane,
            Charing Cross, W.C.
            ...
```

Figure 1. Automatically created Gutenberg knowledge resources object for document 33483 (geosciences collection, LX, excerpt).

The following excerpt (Figure 2) shows a knowledge resources' object automatically created from an entity of Gutenberg document 25062 [21] with LoC classification [25].

```
pg25062.txt [Document]:
            ...
            A STUDY OF
            RECENT EARTHQUAKES.

            BY
            CHARLES DAVISON, Sc.D., F.G.S.

            AUTHOR OF
            "THE␣HEREFORD␣EARTHQUAKE␣OF␣DECEMBER␣17TH,␣1896."
            ...
            WITH 80 ILLUSTRATIONS
            ...
            London and Newcastle-on-Tyne:
            THE WALTER SCOTT PUBLISHING CO., LTD.
            1905
            ...
            PREFACE.
            ...
            The present volume differs from a text-book of seismology in giving
            brief, though detailed, accounts of individual earthquakes rather
             than
            a discussion of the phenomena and distribution of earthquakes in
            general. ...
```

Figure 2. Automatically created Gutenberg knowledge resources object for document 25062 (geosciences collection, LX, excerpt).

Both objects share the same LoC classification. Without advanced means and further analysis both might be considered providing knowledge for the same topics and purposes.

On the other hand, as an example from a different category of resources, an object excerpt of an object instance “Vesuvius” from a knowledge resources’ collection, resulting from a knowledge mining process, is shown in Figure 3.

```
Vesuvius [Volcanology, Geology, Archaeology]:
          (lat.) Mons Vesuvius.
          (ital.) Vesuvio.
          Volcano, Gulf of Naples, Italy.
          Complex volcano (compound volcano).
          Stratovolcano, large cone (Gran Cono).
          Volcano Type: Somma volcano,
          VNUM: 0101-02=,
          Summit Elevation: 1281\UD{m}. ...
          ...
          Syn.: Vesaevus, Vesevus, Vesbius, Vesvius
          s. volcano, super volcano, compound volcano
          s. also Pompeji, Herculaneum, seismology
          ...
          compare La Soufrière, Mt. Scenery, Soufriere
          ...
          %%IML: UDC:[911.2+55]:[57+930.85]:[902]"63"(4+37+23+24)=12=14
          %%IML: GoogleMapsLocation: http://maps.google.de/maps?hl=de&gl=de&vpsrc
           =0&ie=UTF8&ll=40.821961,14.428868&spn=0.018804,0.028238&t=h&z=15
          ...
          ...
          ...
          ...
          ...
          ...
```

Figure 3. Knowledge resources collection object “Vesuvius” (LX resources, geoscientific collection, excerpt).

The objects can contain any knowledge, e.g., factual and conceptual knowledge. Here, the object carries names and synonyms in different languages, dynamically usable geocoordinates, Universal Decimal Classification (UDC) and so on, including geoclassification (UDC:(37), Italia. Ancient Rome and Italy). In addition to collection objects, another important source of knowledge are container type resources, which contain objects, mostly highly comparable from perspectives of content and structure. Figure 4 shows a tiny excerpt of a processed volcanological features container.

```
UCC:UDC2012:551.21
UCC:UDC2012:551
UCC:UDC2012:551.2,551.23,551.24,551.26
UCC:UDC2012:902/908
UCC:MSC2010:86,86A17,86A60
UCC:LCC:QE521-545
UCC:LCC:QE1-996.5
UCC:LCC:QC801-809
UCC:LCC:CC1-960,CB3-482
UCC:PACS2010:91.40.-k
UCC:PACS2010:91.65.-n,91.
UCC:PACS2010:91.40.Ge,91.40.St,91.40.Rs,*91.45.C-,*91.45.D-,90
...
CONTAINER_OBJECT_EN_ITEM: Vesuvius
CONTAINER_OBJECT_DE_ITEM: Vesuv
CONTAINER_OBJECT_EN_PRINT: Vesuvius
CONTAINER_OBJECT_DE_PRINT: Vesuv
CONTAINER_OBJECT_EN_COUNTRY: Italy
CONTAINER_OBJECT_DE_COUNTRY: Italien
CONTAINER_OBJECT_EN_CONTINENT: Europe
CONTAINER_OBJECT_DE_CONTINENT: Europa
CONTAINER_OBJECT_XX_LATITUDE: 40.821N
CONTAINER_OBJECT_XX_LONGITUDE: 14.426E
CONTAINER_OBJECT_XX_HEIGHT_M: 1281
CONTAINER_OBJECT_EN_TYPE: Complexvolcano
CONTAINER_OBJECT_DE_TYPE: Komplex-Vulkan
CONTAINER_OBJECT_XX_VNUM: 0101-02=
```

Figure 4. Knowledge resources container object, processed instance of a simple container entry “Vesuvius” (LX resources, excerpt).

The container objects comprise of various knowledge, especially factual and conceptual references, including concordances. Objects can also contain multi-lingual entries. The organisation of the objects in a specific container is similiar to identical. The resources’ access and processing can be done in any programming language, assuming that the interfaces are implemented. For example, combining scripting, filtering, and parallel programming can provide flexible approaches.

The data used here is based on the content and context from the knowledge resources, provided by the LX Foundation Scientific Resources (LX not an acronym) [4]. The integration and implementation of conceptual knowledge and concordances as shown part of the objects will be discussed in the next section.

IV. Integration of conceptual knowledge

Conceptual knowledge (e.g., classification) is a very important means for describing objects. For the Gutenberg resources a simple LoC classification for every document is provided, which allows to get the major topic classification.

The knowledge resources can use any classification for describing objects, elements, and views. This also allows to use multiple classifications and even specialised classifications for the description. The resources can also provide concordances for mapping the classifications.

All these means can be used for the documentation and analysis of objects, e.g., comparisons. The conceptual knowledge provides the facilities to handle objects and entities even from different sources and context.

A. Implemented references to conceptual knowledge

The objects resulting from the mining request for “Vesuvius” in the context of “volcanology” are also referred with conceptual knowledge. Table I shows some examples of Gutenberg objects and their classification referenced with the knowledge resources.

Table I. Gutenberg conceptual knowledge: LoC.

Ref.	*LoC Code and Description*		*LoC Ref.*
[18]	QE:	Science: Geology	[25]
[19]	DH:	History: General and Eastern Hemisphere: Netherlands, Belgium, Luxemburg	[26]
[20]	QB:	Science: Astronomy	[27]
[21]	QE:	Science: Geology	[25]
[28]	PS:	Language and Literatures: American and Canadian literature	[29]
[22]	QE:	Science: Geology	[25]
[23]	QE:	Science: Geology	[25]
[24]	QE:	Science: Geology	[25]

A means of integration with other classifications is the use of concordances. Table II shows the according excerpt of concordances with LoC and UDC.

Table II. LoC to UDC concordances.

LoC Code	*UDC Code*	*UDC Reference*
QE: . . .	UDC:55	[30]
DH: . . .	UDC:93/94	[31][32][33]
QB: . . .	UDC:52	[34]
QE: . . .	UDC:55	[30]
PS: . . .	UDC:821.111	[35]
QE: . . .	UDC:55	[30]
QE: . . .	UDC:55	[30]
QE: . . .	UDC:55	[30]

Classifications of UDC editions consistently refer to verbal descriptions. For this part of the research all small unsorted excerpts of the knowledge resources objects only refer to main UDC-based classes, which for this part of the publication are taken from the Multilingual Universal Decimal Classification Summary (UDCC Publication No. 088) [36] released by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license [37] (first release 2009, subsequent update 2012). Table III lists the UDC classifications and their verbal description in English.

TABLE III. UDC RESOLVED, VERBAL DESCRIPTION ENGLISH.

UDC Code	*Description*	*Ref.*
UDC:55	Earth Sciences. Geological sciences	[30]
UDC:93/94	History	[31][32][33]
UDC:52	Astronomy. Astrophysics. Space research. Geodesy	[34]
UDC:55	Earth Sciences. Geological sciences	[30]
UDC:821.111	English literature	[35]
UDC:55	Earth Sciences. Geological sciences	[30]
UDC:55	Earth Sciences. Geological sciences	[30]
UDC:55	Earth Sciences. Geological sciences	[30]

The table lists those entries resulting from the concordances and from associated objects. The UDC is available in about 50 languages. The English verbal descriptions were used for this case study. The verbal description can be included in the context creation and can, for example, provide scalable fuzziness for creating multi-disciplinary context. The LX knowledge resources' structure and the classification references [38] based on UDC [39], [40], [36] are essential means for the processing workflows and evaluation of the knowledge objects and containers. Both provide strong multi-disciplinary and multilingual support.

B. Concordances

Different resources as well as different disciplines may use different classifications. Gutenberg is currently using the LoC. The knowledge resources are using a Universal Classified Classification (UCC) both with classification and concordances for the objects collected over time. The listing in Figure 5 shows a simple example for concordances.

```
...
UCC:UDC:55
UCC:LCC:QE

...
UCC:UDC:93/94
UCC:LCC:DH

...
UCC:UDC:52
UCC:LCC:QB

...
UCC:UDC:821.111
UCC:LCC:PS

...
```

Figure 5. Classification and concordances excerpt of a simple object instance (knowledge resources collection).

The references to individual, specialised and universal classifications consistently describe the conceptual knowledge as correct as possible within a classification.

The concordances integrate the context of more than one classification. This enhances facilities for in-depth description and integration, the specialisation on conceptual knowledge as well as the broadness of universal knowledge and context. The listing in Figure 6 shows a simple object instance classification and concordances excerpt from a volcanological object in a collection.

```
...
UCC:UDC2012:551.21
UCC:UDC2012:551
UCC:UDC2012:902/908
UCC:MSC2010:86,86A17,86A60
UCC:LCC:QE521-545
UCC:LCC:QE1-996.5
UCC:LCC:QC801-809
UCC:LCC:CC1-960,CB3-482
UCC:PACS2010:91.40.-k
UCC:PACS2010:91.65.-n,91.
```

Figure 6. Classification and concordances excerpt of a simple object instance (knowledge resources collection).

The excerpt shows classification concordances in several different classifications as used in different disciplines. Even a lot of internal details of such concordances are self-explanatory with a basic knowledge and practice of the used classifications. Concordances also interlink different classifications and disciplines. To a certain extent most classifications also express the context of certain disciplines. Possibly multiple views from different disciplines or author groups on a certain object are not shown in this reduced view but they can also hold the full spectrum of classifications and concordances and also express views and development of views and object instances over time.

V. KNOWLEDGE-CENTRIC INTEGRATIVE ARCHITECTURE

The analysis of integrated resources requires advanced methods and algorithms. A method used for description and analysis of objects is the Content Factor.

A. Content Factor computation for data entities

Objects of any kind can be integrated with knowledge resources. Objects can contain instances of data entities and refer to associated knowledge. For an analysis, a number of common information regarding the objects and data entities is required.

Examples were shown in the previous section on object and data entity integration, where objects were transformed into collection objects, including references to further knowledge like factual knowledge and classification.

The following excerpt (Figure 7) illustrates the creation of Content Factor definition sets [2] for the use with data entities. All the basics details and the algorithm of the method are described in the original publication. Definition sets are used for both Gutenberg resources and knowledge resources.

```
% (c) LX-Project, 2016, 2017
{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
{Veu}:=[Vv][Ee][Ss][Uu][Vv]
{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
{Kom}:=[Kk][Oo][Mm][Ee][Tt]
{Com}:=[Cc][Oo][Mm][Ee][Tt]
{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
```

Figure 7. CONTFACT definition set for Gutenberg Project resources and knowledge resources, (LX, excerpt).

Figure 8 shows the Normed Basic Content Factor (NBCF, $\overline{\kappa}_{\mathrm{B}}$) [2] computed for a knowledge resources object reference to the Gutenberg Project document 33483.

```
CONTFACT:BEGIN
CONTFACT:20161227-234624:AU:{Veu}{Veu}{Vul}{Vol}{Ear}{Veu}{Met}{Veu}{Vol}{Met
 }...{Veu}{Veu}{Ear}{Veu}...{Veu}{Veu}{Veu}{Veu}...{Veu}{Ear}{Vol}{Ear}/39843
CONTFACT:20161227-234624:AS:{Ear}...{Veu}{Veu}{Vol}{Vol}...{Vul}{Vul}/39843
CONTFACT:20161227-234624:M:{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234624:M:{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234624:M:{Veu}:=[Vv][Ee][Ss][Uu][Vv]
CONTFACT:20161227-234624:M:{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
CONTFACT:20161227-234624:M:{Kom}:=[Kk][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234624:M:{Com}:=[Cc][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234624:M:{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
CONTFACT:20161227-234624:M:{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
CONTFACT:20161227-234624:M:{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
CONTFACT:20161227-234624:M:{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234624:M:{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSDEF=11
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSALL=39843
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSMAT=356
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSCFO=.00900324
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSKWO=1
CONTFACT:20161227-234624:M:STAT:OBJECTELEMENTSLAN=0
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSOBJ=33483-0.txt
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSDCM=(c) LX-Project, 2016, 2017
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSMTX=LX Foundation Scientific
 Resources; Object Collection
CONTFACT:20161227-234624:M:INFO:OBJECTELEMENTSAUT=Claus-Peter R\"uckemann
CONTFACT:END
```

Figure 8. NBCF $\overline{\kappa}_{\mathrm{B}}$ computed for knowledge resources object reference to Gutenberg Project document 33483 (LX Resources, excerpt).

Figure 9 shows the Normed Basic Content Factor (NBCF, $\overline{\kappa}_{\mathrm{B}}$) computed for a knowledge resources object reference to the Gutenberg Project document 25062.

```
CONTFACT:BEGIN
CONTFACT:20161227-234626:AU:{Ear}...{Vol}{Ear}...{Veu}{Vol}{Ear}...{Veu}{Ear}{
 Ear}{Ear}{Veu}...{Veu}...{Veu}{Ear}...{Ear}/88463
CONTFACT:20161227-234626:AS:{Ear}...{Veu}...{Vol}{Vul}{Vul}/88463
CONTFACT:20161227-234626:M:{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234626:M:{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
CONTFACT:20161227-234626:M:{Veu}:=[Vv][Ee][Ss][Uu][Vv]
CONTFACT:20161227-234626:M:{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
CONTFACT:20161227-234626:M:{Kom}:=[Kk][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234626:M:{Com}:=[Cc][Oo][Mm][Ee][Tt]
CONTFACT:20161227-234626:M:{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
CONTFACT:20161227-234626:M:{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
CONTFACT:20161227-234626:M:{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
CONTFACT:20161227-234626:M:{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234626:M:{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSDEF=11
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSALL=88463
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSMAT=986
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSCFO=.01122068
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSKWO=1
CONTFACT:20161227-234626:M:STAT:OBJECTELEMENTSLAN=0
CONTFACT:20161227-234626:M:INFO:OBJECTELEMENTSOBJ=pg25062.txt
CONTFACT:20161227-234626:M:INFO:OBJECTELEMENTSDCM=(c) LX-Project, 2016, 2017
CONTFACT:20161227-234626:M:INFO:OBJECTELEMENTSMTX=LX Foundation Scientific
 Resources; Object Collection
CONTFACT:20161227-234626:M:INFO:OBJECTELEMENTSAUT=Claus-Peter R\"uckemann
CONTFACT:END
```

Figure 9. NBCF $\overline{\kappa}_{\mathrm{B}}$ computed for knowledge resources object reference to Gutenberg Project document 25062 (LX Resources, excerpt).

Figure 10 shows the NBCF computed for a knowledge resources object reference to the object “Vesuvius” (Figure 3).

```
CONTFACT:BEGIN
CONTFACT:20170205-161508:AU:{Veu}{Vol}{Veu}{Veu}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{
 Vol}{Vol}{Vol}/71
CONTFACT:20170205-161508:AS:{Veu}{Veu}{Veu}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{Vol}{
 Vol}{Vol}{Vol}/71
CONTFACT:20170205-161508:M:{Vol}:=[Vv][Oo][Ll][Cc][Aa][Nn]
CONTFACT:20170205-161508:M:{Vul}:=[Vv][Uu][Ll][Cc][Aa][Nn]
CONTFACT:20170205-161508:M:{Veu}:=[Vv][Ee][Ss][Uu][Vv]
CONTFACT:20170205-161508:M:{Vee}:=[Vv][Ee][Ss][Ee][Vv][Oo]
CONTFACT:20170205-161508:M:{Kom}:=[Kk][Oo][Mm][Ee][Tt]
CONTFACT:20170205-161508:M:{Com}:=[Cc][Oo][Mm][Ee][Tt]
CONTFACT:20170205-161508:M:{Met}:=[Mm][Ee][Tt][Ee][Oo][Rr]
CONTFACT:20170205-161508:M:{Erd}:=[Ee][Rr][Dd][Bb][Ee][Bb][Ee][Nn]
CONTFACT:20170205-161508:M:{Ear}:=[Ee][Aa][rr][Tt][Hh][Qq][Uu][Aa][Kk][Ee]
CONTFACT:20170205-161508:M:{Puz}:=[Pp][Uu][Zz][Zz][Oo][Ll]
CONTFACT:20170205-161508:M:{Poz}:=[Pp][Oo][Zz][Zz][Oo][Ll]
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSDEF=11
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSALL=71
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSMAT=13
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSCFO=.21311475
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSKWO=2
CONTFACT:20170205-161508:M:STAT:OBJECTELEMENTSLAN=1
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSOBJ=Vesuvius
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSDCM=(c) LX-Project, 2016, 2017
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSMTX=LX Foundation Scientific
 Resources; Object Collection
CONTFACT:20170205-161508:M:INFO:OBJECTELEMENTSAUT=Claus-Peter R\"uckemann
CONTFACT:END
```

Figure 10. NBCF $\overline{\kappa}_{\mathrm{B}}$ computed for knowledge resources object “Vesuvius” (LX Resources, excerpt).

All NBCF were computed with the same definition set (Figure 7). The data entities from the referenced Gutenberg resources and knowledge resources both contain multiple matches. The resulting Content Factor for the knowledge resources object is higher due to the higher concentration of relevant elements in the object. The Gutenberg object shows a higher absolute number of matches and multiple hits.

B. *Architecture and integration*

The Content Factor provides a measure, which can be used for describing characteristics of different objects.

The method can be used in general for any kind of object and is not restricted to limited kinds of documents. The Gutenberg document resources were choosen for the case study because of their range of content, type of documents, and availability of different text formats, which correspond with the intended integration with other resources.

The following diagram (Figure 11) depicts the architecture design of the conceptual solution as used for the implementation. The illustration shows the categories of resources and their integration regarding their contributions to a Comparative Analysis workflow, implemented for the use with the Gutenberg resources.

The workflow shows the abstract mathematical and computations details of the integration of resources, especially the interfaces, analysis, join, visualisation, statistics, and plotting as of the developed solution. With this research, resources can be divided in three major categories, symbolised as columns, depending on their nature and characteristics.

The central resources are knowledge resources, containing collections and containers as well as referenced and integrated resources. Content is arbitrary, e.g., any factual knowledge, conceptual knowledge, and procedural knowledge.

In this case, the enhancement of the means of mining and the knowledge resources are subject originary resources, especially the resources of the Gutenberg Project [5].

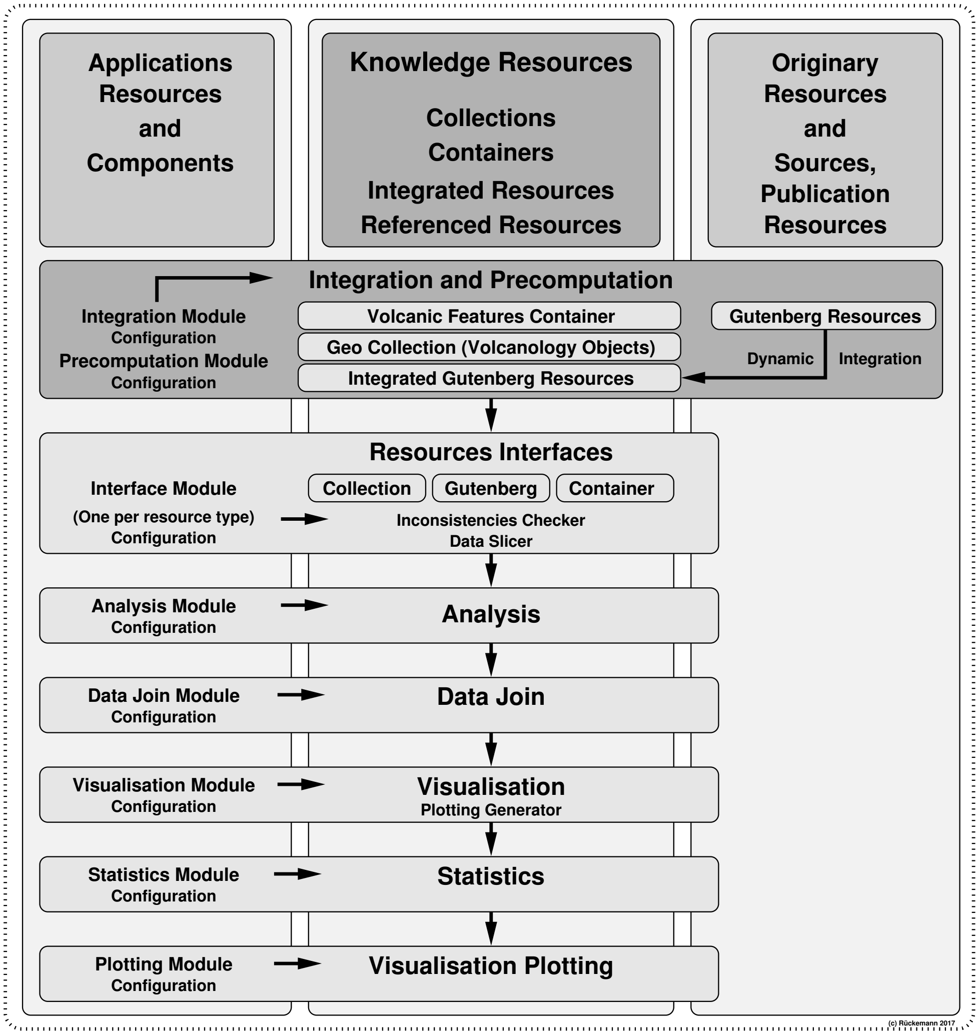

Figure 11. Architecture and integration: Categories of resources and their integration, showing the contributions to a Comparative Analysis workflow, implemented for the use with the Gutenberg resources being dynamically integrated, e.g., for analysis, visualisation, and statistics in an automated decision making workflow.

For the mining example, the integration and precomputation has to care for volcanic features container, geo collection, and the integrated objects from the Gutenberg resources.

C. Knowledge Resources and originary resources

The analysis of different classifications, development of concepts for intermediate classifications, and experiences from case studies from the research conducted in the Knowledge in Motion (KiM) long-term project [41] have contributed to the application of UDC and different classifications and concordance schemes in the context of knowledge resources.

The following term definitions for object, container, and matrix can be helpful in this context.

- An object is an entity of knowledge data being part of knowledge resources. An object can contain any documentation, references, and other data. Objects can have an arbitrary number of sub-objects.
- A container is a collection of knowledge objects in a conjoint format.
- A matrix is a subset of the entirety, the “universe”, of knowledge. A workflow can consist of many sub-workflows each of which can be based on an arbitrary number of knowledge matrices. The output of any sub-workflow or workflow can be seen as an intermediate or final result matrix.

The flexible creation of objects carrying references, especially classification and concordances is the fundament for advanced. knowledge processing and computing.

Collections can hold objects, which are more or less individual consist of any smaller entities and can have any references to other objects or resources. Containers can hold object, which have a comparable structure and comparable entities, e.g., objects belonging to a certain field of research or discipline. External resources can be referred from these objects and resources but it is also possible to integrate objects, collections, and containers from external resources.

The category of originary resources can contain realia objects, original digital objects, as well as sources and publication resources, e.g., physical or digital books or proceedings. Most of the resources in this category restrict access to their content and provide limited interfaces and individual structures.

D. Computation and integration

The category of application resources and components can hold applicable implementations, e.g., software routines, interfaces, and services.

The Gutenberg resources were choosen for this implementation and the practical case study. The Gutenberg resources provide published materials for free Open Access, especially books, but neither provide interfaces nor further usable structures like containers in the aforementioned meaning.

The implementation therefore considered an integration of Gutenberg resource entities into the knowledge resources, on the level of collections and containers. The integration can be done dynamically and non-dynamically as well as the integration can be persistent with the knowledge resources or not. In any case, the reason is to created objects and entities, which are at a comparable level with other available objects, in content type and structure. The objects and entities are precomputed for that purpose. Therefore, with the application resources, the implementation can provide modules for the integration and precomputation. In consequence, the workflow (Figure 11) can rely on the integrated resources and the modules provided by the application resources.

The workflow requires interface modules for the participating resources, one individual module per resource type. This step includes additional preparation, e.g., inconsistencies checkers and data slicers. Data slicers are partitioning the available content in a way in which is seems reasonable to have a common analysis with the integrated objects, e.g., based on full objects, entities, text blocks or lines.

The next steps do the analysis of objects, also considering their individual nature and original, in order to prepare for joining the data. A consequent step does an intermediate visualisation, generating a plotting routine for further consequent statistics, analysis and visualisation. The plotting generator is especially significant in this case study because the partial intermediate visualisation can deliver important information for further statistics and visualisation.

VI. Implementation of the workflow

A. Procedures and modules

Two main modules were required with the assistance precomputation for identifying and selecting objects and data entities from the Gutenberg resources before entering the CA workflow. The preparative assistance data was computed with a module `gutenberganalysis` and the classification was extracted with a module `gutenbergloc`.

The first module extracts the desired content of an specific object and checks for association with the mining request, e.g., keywords and referenced knowledge. The second module extracts the classification of a specific object.

The CA workflow builds on these preparatory results. The implementation case study for the comparative analysis method required the creation of several major components and modules. Table IV shows a sequence of modules, which allows to create the base for a CA workflow as created with this case study.

Table IV. Comparative Analysis workflow procedures and implemented modules with Gutenberg resources.

Procedure	*Module*
Gutenberg interface	`textca_gutenberginterface`
Configuration	
Inconsistencies checker	
Data slicer	
Analysis	`textca_analysis`
Configuration	
Data join	`textca_join`
Configuration	
Visualisation module	`textca_visualisation`
Configuration	
Plotting generator	
Conditional visualisation	
Statistics	`textca_statistics`
Configuration	
Visualisation plotting	`textca_plotting`
Configuration	

The workflow shows the principle procedures. Practically, the modules can be implemented with any environment and frameworks. In the case study Perl [42], Shell, and Gnuplot [43] were used. In general this means any module could be replaced by a different implementation separately.

Any module requires configuration options, which at least can be pre-configured options. In their application, the analysis up to visualisation modules for the Gutenberg resources are identical to the application for the knowledge resources. Therefore, the computations for all data entities were done with the `textca` group of modules. The module special for the Gutenberg resources is the Gutenberg interface, which requires to implement the proper handling as desired for structure and content of the integrated knowledge.

The next sections will show how a CA analysis is done in detail and discusses some examples of possible characteristics, which can be read from the results of the analysis. The analysis is done for objects and resources, which were already introduced. The examples include knowledge collections and containers. Example outputs of modules are given as is, meaning the figures show direct visualisation output from the implementation.

B. Comparison of data entities in and with collections

The following part illustrates characteristics of an integrated object and a knowledge resources collection object. Figure 12 shows the automatically computed CA module result for a case insensitive `vesuv` (`[Vv][Ee][Ss][Uu][Vv]`) target for the above Gutenberg object instance (Figure 1) of the originary object [19].

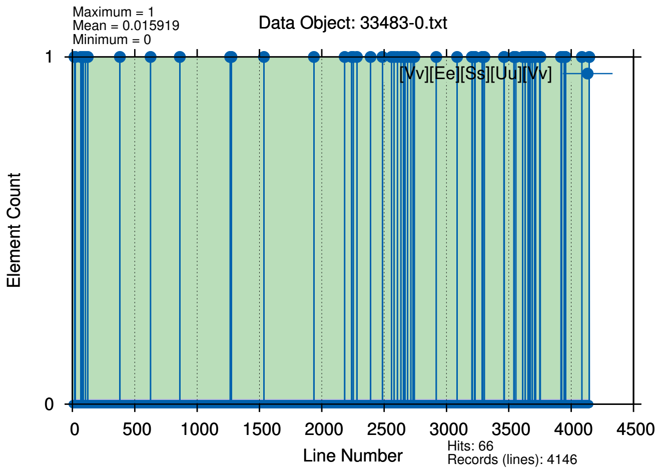

Figure 12. Comparative Analysis module result for a Gutenberg object precomputed by an assistance process for the case insensitive `vesuv` target.

The analysis including the illustration was automatically computed for the respective object. The results are shown on the background of a representation of the object entity. Resulting element counts displayed against line numbers reveal some respective characteristics of the object entity. The object entity extends over the maximum count / line number range (greenish color). Some assistive values are given, like number of records (lines in this case), number of hits, maximum, minimum, and mean value of hits. Despite the large number of hits, here, the mean value (blueish color) of hits is quite low, due to the relatively large number of records. The result also reveals sequences of higher hit-density and hit-patterns in the object entity in the illustration.

Figure 13 shows the automatically computed CA module result with the respective target (pattern) for the resulting knowledge resources collection object “Vesuvius” (Figure 3).

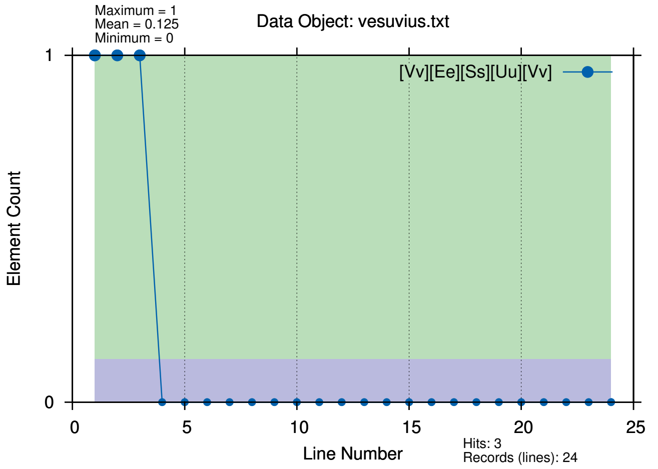

Figure 13. Comparative Analysis module result for the knowledge resources collection object “Vesuvius” (LX resources, geoscientific collection, excerpt).

The result shows some criteria of the object itself in context with the relevant mining pattern. The figure illustrates that the object contains a relevant mining result in the first and several consecutive records (here: lines) with a maximal occurance count of one in a record.

The density of relevant occurances in the object is relatively high compared to common texts, even if from comparable special topic documents. Therefore, the mean value is quite high in that case. The computed background shading illustrates the space spanned by the available records (number of lines) and element counts. The mean value is illustrated by the border of the color change.

Both figures show that all terms were considered case insensitive. Here, collection objects and integrated object documents are choosen, which are referring to Vesuvius. For both, the mining workflow considers different ways of writing. Thus, for example, associations of Vesuvius and Pozzulan as well as directly linking to volcanic features and meteoric features via different resources.

The differences are intrinsic characteristics of the two types of objects, e.g., the length of the object, the concentration of hits, especially at the top of the object, and relatively high mean value of hits. The characteristics can be used to automatically decide to which extend objects can contribute to enhancements. Longer objects can contain several passages of records, which can be compared separately with an object and contribute to the enhancement of resources.

Features and properties of objects in collections are different to objects in containers. In many cases the differences of general objects in the Gutenberg resources can best be described in collections. Nevertheless, the comparison of collection and container objects will reveal characteristics of objects in both, which can contribute to the enhancement of workflow results.

C. Comparison of data entities in and with containers

Figure 14 shows the computed CA module result for a Gutenberg object instance (Figure 9, originary object [21]) for a case insensitive `vulc/volc` (`[Vv][UuOo][Ll][Cc]`) target.

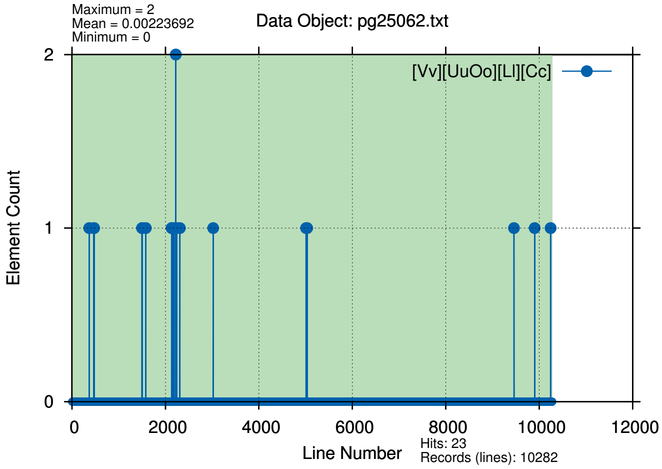

Figure 14. Comparative Analysis module result for a Gutenberg object precomputed by an assistance process for case insensitive `vulc`/`volc` target.

Figure 15 shows the automatically computed CA module result with the respective target (pattern) for the resulting knowledge resources collection object “Vesuvius” (Figure 3).

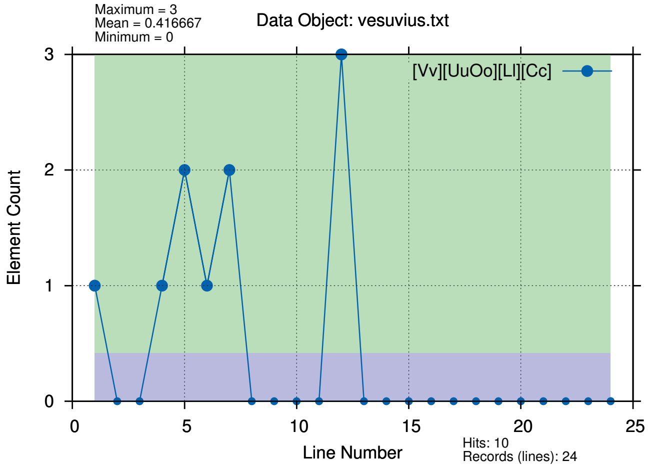

Figure 15. Comparative Analysis module result for the knowledge resources collection object "Vesuvius" (LX resources, geoscientific collection, excerpt).

There is more than one occurance in several lines each, with a maximal occurance count of three in a record. Figure 16 shows the computed CA module result for the volcanological features container (Figure 4) for the same target.

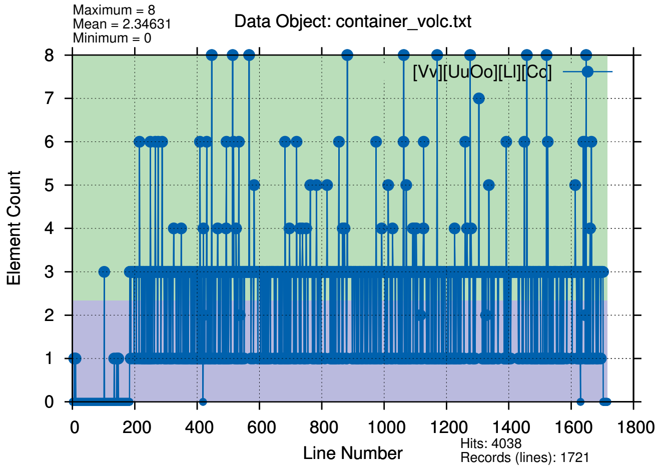

Figure 16. Comparative Analysis module result for the volcanological features container for case insensitive `vulc`/`volc` target (LX resources).

Both figures (Figures 16 and 14) illustrate the very high relevance of the objects. Nevertheless, the structure and density of hits is much higher in the container object than in the Gutenberg object. In addition, the mean value is extremely high for the container object. Also, the central part of the container object does not contain a line without the target. There are even more hits than records.

Even in a top hit Gutenberg object the number of records is much higher and the number of hits is lower. The comparison also reveals that both objects represent different object types, a knowledge resources object and a classical text object. The latter one mostly contains natural language. For any resources, many CA and Content Factor computations are done on a result matrix. With any workflow further information and decision making support can result from computing assistant views for the knowledge entities, e.g., based on their context.

D. *Assistant views*

For any type of resources, many CA and Content Factor computations are done on a result matrix. Besides the knowledge mining request "Vesuvius" in the context of "volcanology" with this example, the comparison with other requests can be helpful.

For example, common words, e.g.,'the', which does rarely occur in containers but mostly in book texts can be used for automated decision making. Figure 17 shows the computed CA module result for a Gutenberg object for a very common target, `the`.

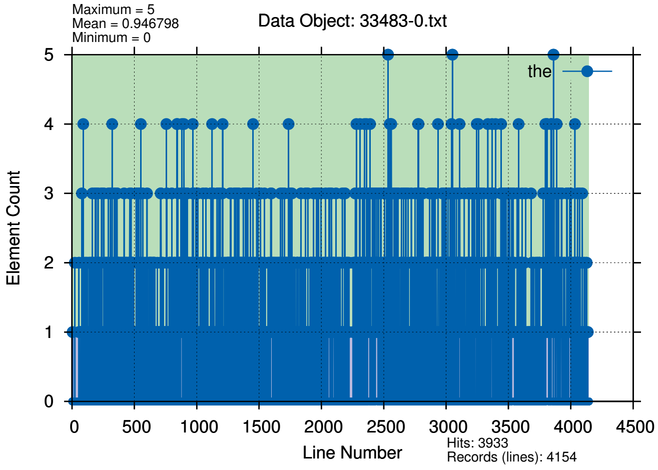

Figure 17. Comparative Analysis module result for a Gutenberg object precomputed by an assistance process for the case insensitive `the` target.

In result, the relative density is an excellent indicator for a longer natural language object. The accumulation of hits, e.g., a wrapping curve, indicates a longer homogeneously structured natural language object. The distribution indicates that the respective natural language document rarely contains longer passages of content, which will represent container-like knowledge. The results of the CA, the Content Factor, classification, and any results and attributes from assistant views can be included in an workflow and analysis, e.g., if a ranking of results is required for a specific knowledge mining workflow.

VII. Discussion

The case study integrates sources of different knowledge entities for knowledge mining workflows, selecting entities by computing advanced analysis criteria. The combination of advanced methods and the integration of resources are the essential basis for enhancement processes.

A. *Integration and comparison*

The selected data compasses over 50,000 Gutenberg documents and more than 50,000 objects from knowledge resources. The selected sizes of objects range from hundreds of bytes to several megabytes.

The limitation for the case study was done for demonstration, due to the fact that the number of available overall knowledge resources objects may easily outnumber the number of Gutenberg documents. With the resources, the classification considers about fifty languages, summing up to about three million descriptions. Conceptual assistance is available for resource and object classification, which allows to automate integration workflows.

For the integration, instances of the objects containing the relevant data entities were automatically computed. It was possible to apply the provided means in the same way to the entities. The computation for the data entities from knowledge resources can be much more fine grained and systematic due to the complex structures and elements. The computation for the Gutenberg data entities can use the same means but some details and structure are not automatically available. The data sizes of the main Gutenberg data entities are most probably larger than these of the average knowledge resources' data entities.

The Content Factor method delivers an efficient tool in order to select object entities for their later integration with the knowledge resources and discovery workflows. The resulting objects can be easily CA computed and analysed regarding their contributions to enhancing the resources as well as the results of discovery.

Terms like 'precision' to not apply when dealing with complex knowledge, especially when knowledge entities are containing or are described by natural language content. In these terms, 'precision' depends on the question asked by the person implementing a scenario. Here, e.g., the CA mean values fill in for the purpose of a suitable measure. The range is from maximum fit to minimum fit.

B. CA mean values

Table V compares CA mean values from the computation for selected objects and target groups for the integrated resources.

TABLE V. SELECTED COMPUTATION DATA ENTITIES: OBJECTS AND TARGET GROUPS SORTED BY THEIR CA MEAN VALUES.

Object	*Target / Target-Group*	*CA Mean*
Knowledge res., Vesuvius	`[Vv][Ee][Ss][Uu][Vv]`	0.125
Gutenberg 33483-0	`[Vv][Ee][Ss][Uu][Vv]`	0.015919
Know. res., volc. feat. cont.	`[Vv][Ee][Ss][Uu][Vv]`	0.00291375
Gutenberg 25062	`[Vv][Ee][Ss][Uu][Vv]`	0.000486287
Know. res., volc. feat. cont.	`[Vv][UuOo][Ll][Cc]`	2.34631
Knowledge res., Vesuvius	`[Vv][UuOo][Ll][Cc]`	0.416667
Gutenberg 33483-0	`[Vv][UuOo][Ll][Cc]`	0.0356971
Gutenberg 25062	`[Vv][UuOo][Ll][Cc]`	0.00223692

There are entities with higher and lower mean values, for the Gutenberg resources as well as for the knowledge resources. Higher values indicate a cumulation of relevant terms, e.g., as with the appearance in collections, tabulars, and listings.

Practice showed that for complementing knowledge in the volcanological features container with extended context, relevant entities from the Gutenberg resources with higher mean values can be a primary source for references. Relevant entities from the Gutenberg resources with lower mean values may primarily deliver reference information for collection objects.

C. Comparisons in computation and analysis

As illustrated, longer objects can contain several passages of records, which can be compared separately with an object and contribute to the enhancement of resources. Beyond that, decisive workflows can benefit from the comparative results as to choose the best-fit objects for decisions, especially selecting the most fitting/associated passages in a text and comparing it to the most fitting/associated passage in another text.

Decisive workflows of that dimension are very challenging regarding computation and analysis. Table VI lists the counts of comparisons for the application in the above case study. The values are given for a single mining request (Figure 7).

TABLE VI. REPRESENTATIVE COUNTS OF COMPARISONS PER MINING REQUEST WITH THE CASE STUDY OBJECT ENTITIES.

Comparisons	*Count*
entities in collections	720,000
entities in containers	21,000
integrated documents, Gutenberg subset	570,000
overall, within filename/label space	550,000
overall, first result level	35,000
overall, second result level	15,000
overall, in pre-final results	4,500

A number of comparisons have to be done per mining request, depending on the complexity of the request and the configuration of the resources. In the above case, collections, containers, and integrated resources were used. Counts of comparisons were done in the respective entities and content, not in references outside the subsets.

In consequence, comparisons were done over all of the configured resources. The comparisons also included the filenames of the known integrated objects and the labels of the collection and containers entities because these regulary contain relevant knowledge for mining requests.

Most mining requests benefit from checking the result and trying to get additional information and enhanced results from the first results, using the same resources. In that case the first, second, and further result levels are considered intermediate results, which are used to go in depth and width for the consecutive mining.

The counts illustrate that a mining request can be very challenging, even for a subset of configured resources. There can also be needs to pre-cache and pre-compute resources, depending on the purpose of the application scenario in which the mining request should be embedded. In this case, the levels of the intermediate result generation can be considered an enhancement of selection, which leads to less comparisons required with the levels.

D. Ranking

For this scenario, a ranking was built from the rankings for the entities from the Gutenberg entities and from the knowledge resources.

The ranking considers the available information, e.g., classified targets, relevance of targets, references and context. The Gutenberg ranking especially considers the results from the CA, Content Factor and classification (LoC), based on the primary Gutenberg result matrix. The ranking of knowledge resources especially considers the CA, Content Factor, and classification, e.g., UDC and Universal Classified Classification (UCC). The integrated ranking considers the the CA, Content Factor values, and concordances of comparable entities.

An integration for a workflow ranking requires that the means need to be individually choosen for a certain application scenario. In this case, a records base (lines) was an appropriate choice for CA, Content Factor, and conceptual knowledge.

E. Computational trace and context

A common knowledge discovery process integrates a sequence of decision making processes at different levels, e.g., from which resources to which single objects. Each step in a sequence can require to handle millions of objects and references. The access to the Gutenberg resources is not intended to be automated. Therefore, no performance data is available for the Gutenberg resources or for conducting the precomputation for its whole content. The precomputation assistance includes the cached Gutenberg content for the respective mining targets.

Table VII shows the computation characteristics relevant with the workflow procedure for an example of the above integrated Gutenberg and knowledge resources case for two objects.

TABLE VII. COMPUTATION CHARACTERISTICS WITH THE WORKFLOW PROCEDURE FOR TWO INTEGRATED OBJECTS, WALL TIMES PER CPU.

Workflow Procedure	*Wall Time*
Precomputation assistance	24.8 s
Analysis, resources classification	1.2 s
Analysis, object classification	14.7 s
Comparative Analysis	3.2 s
(Integrative workflow step)	n s

The table times refer to one Central Processing Unit (CPU) per mining process (Intel Xeon, at 2.9 GHz). Due to the complexity of the elementary workflows it is not desirable to have more than one CPU per process involved at the atomic level. Arbitrary practical application scenarios involving many processes with large data resources may be organised to fit the architecture of the available infrastructure. With certain scenarios, where an author wants to integrate complex references, the precomputation assistance can benefit a lot from using many-CPU infrastructures.

The higher level workflow step, integrating the aforementioned procedures, will use a lot of intermediate results from procedures and content from resources. There is no general range for the time scale at the higher levels but at these levels the requirements on computation and communication can be extremely high, therefore, the higher level steps are candidates for parallelisation. Anyhow, workflow creators must always be aware that computing requirements can be non-linear, depending on the workflow created by an author for a choosen purpose.

VIII. CONCLUSION

The paper presented the results of a research based on an advanced method for knowledge mining with multi-disciplinary knowledge resources and different data entities. Required modules and algorithms were successfully and efficiently implemented for supporting a Comparative Analysis, integrating different data entities in mining workflows.

This research showed that important characteristics of objects can be automatically identified and the results can be used for enhancing knowledge resources themselves as well as the discovery processes. The basis of the success is that with the availability of appropriate methodologies and methods different entities neither need to be left out from advanced knowledge mining workflows nor should their content be ignored. It was shown that in result, there is a complementary relationship between objects from knowledge resources and referred objects from external sources, including their data entities.

The Content Factor method for data description and analysis is used with all available resources. As was shown, CA methods cannot be replaced by other means like classification or Content Factor because they are based on completely different grounds but these complementary means can be integrated within more complex workflows. Classification is an integral complement in all parts of CA methods. In an implementation it means, e.g., that UDC can be used in all steps and components, starting from the knowledge resources. CA modules can help optimise the decision making, e.g., with supporting context-spanning Content Factor definition sets. CA modules can be used for delivering additional descriptive information, which can be used for documentation and knowledge mining purposes. CA is much beyond statistics. The significant part of the CA is the visualisation of pattern sequences in entities. The pattern sequences hold relevant parts of the entity characteristics and can also be used for documentation. The statistics are used in addition, for the analysis.

In consequence, the method allows an effective integration of resources and a dynamical selection of objects and entities. Selected objects and entities can be used dynamically, e.g., in discovery and decision making but also for persistently enhancing the context and references of available knowledge resources.

Objects from advanced knowledge resources can provide an excellent data base on knowledge. The knowledge resources can provide high quality object collections and containers with data entities of most reliable and unique content and qualities. Referred objects from external sources can extend the available data base regarding width and depth. Therefore, referred objects and external sources can extend the available data base and content. In the case study, the best fit targets regarding volcanological features from the resources, including the Gutenberg resources, were automatically analysed.

The method can enhance the quantity and quality of knowledge resources and dicovery results by integration and analysis of content and context for many application scenarios. In the case studies the quality was verified manually. A single measure of quality is out of scope of this research because thr evaluation would depend on a certain purpose of application.

On the side of the Gutenberg resources, a number of challenges have been found especially with the Gutenberg objects themselves. With the documents, workflow creators face a lot of inconsistencies in structure and marking even regarding major elements. Bibliographic data and versioning could also be improved. Better structured and more complete bibliographic data would be beneficial for any wider and systematic use. A common container format for the Gutenberg documents, handling any data files and associated data in a flexible and 'clean' way would be beneficial.

The integration of the Content Factor method and the Comparative Analysis method is used in practice for progressive advancement of long-term multi-disciplinary knowledge resources and mining in a number of institutions and projects, including the knowledge resources used here for demonstration purposes. Special targets in practice are quality and quantity of entities but also balancing the content and context development of knowledge resources.

Information sciences, especially knowledge mining methods, have countless areas of implementation. Anyhow, the application scenarios of the methodology and the implemented methods are even not limited to knowledge mining. Besides the purpose laid out with this research, CA modules can be used as complementary and supportive methods applied with a wide range of advanced applications like document identification or plagiarism detection. For complex solutions, it can be desirable to integrate many components and modules. From the view of (the modern understanding of) technics, licensing acts as starting point for realisation of technical solutions. In this respect, licensing solutions are an integral part of almost any technical realisation and vice versa. Future work will be spent on further integrating different resources and creating methodologies, methods, and means for handling data entities and objects as well as on realisation aspects and education for the context of superordinate knowledge.

Acknowledgements

We are grateful to the "Knowledge in Motion" (KiM) long-term project, Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF), for partially funding this implementation, case study, and publication under grants D2014F1P04518, D2014F2P04518, D2016F1P04683, and to its senior scientific members, especially to Dr. Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, Germany, to Dipl.-Biol. Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany, to Dipl.-Ing. Martin Hofmeister, Hannover, Germany, and to Olaf Lau, Hannover, Germany, for fruitful discussion, inspiration, practical multi-disciplinary case studies, and the analysis of advanced concepts. We are grateful to Dipl.-Ing. Hans-Günther Müller, Cray, Germany, for his work and assistance providing practical solutions to architectural challenges, practical private cloud and storage solutions, and excellent technical support. We are grateful to all national and international partners in the Geo Exploration and Information cooperations for their constructive and trans-disciplinary support. We thank the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research since 1997, including the GEXI developments and case studies.

References

[1] C.-P. Rückemann, "Comparative Analysis of Data Entities: Knowledge Mining Objects," in Proceedings of The Seventh International Conference on Advanced Communications and Computation (INFOCOMP 2017), June 25–29, 2017, Venice, Italy. XPS Press, Wilmington, Delaware, USA, 2017, Rückemann, C.-P., Flood, I. and Schwerdtfeger, I. and Simmendinger, C. and Beckett, G. (eds.), pages 17–23, ISSN: 2308-3484, ISBN-13: 978-1-61208-567-8, URL: http://www.thinkmind.org/index.php?view=article&articleid=infocomp_2017_3_10_60019 [accessed: 2018-05-10].

[2] C.-P. Rückemann, "Enhancement of Knowledge Resources and Discovery by Computation of Content Factors," in Proceedings of The Sixth International Conference on Advanced Communications and Computation (INFOCOMP 2016), May 22–26, 2016, Valencia, Spain. XPS Press, 2016, pages 24–31, ISSN: 2308-3484, ISBN-13: 978-1-61208-478-7, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2016_2_30_60047 [accessed: 2018-05-10].

[3] C.-P. Rückemann, Z. Kovacheva, L. Schubert, I. Lishchuk, B. Gersbeck-Schierholz, and F. Hülsmann, Best Practice and Definitions of Data-centric and Big Data – Science, Society, Law, Industry, and Engineering. Post-Summit Results, Delegates' Summit: Best Practice and Definitions of Data-centric and Big Data – Science, Society, Law, Industry, and Engineering, Sep. 19, 2016, The Sixth Symp. on Advanced Computation and Information in Natural and Applied Sciences (SACINAS), The 14th Int. Conf. of Numerical Analysis and Applied Mathematics (ICNAAM), Sep. 19–25, 2016, Rhodes, Greece, 2016, URL: http://www.user.uni-hannover.de/cpr/x/publ/2016/delegatessummit2016/rueckemann_icnaam2016_summit_summary.pdf [accessed: 2018-05-10].

[4] "LX-Project," 2018, URL: http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX [accessed: 2018-05-12].

[5] "Project Gutenberg," 2018, URL: http://www.gutenberg.org [accessed: 2018-02-04].

[6] Aristotle, Nicomachean Ethics, 2008, (Written 350 B.C.E.), Translated by W. D. Ross, Provided by The Internet Classics Archive, URL: http://classics.mit.edu/Aristotle/nicomachaen.html [accessed: 2018-05-10].

[7] Aristotele, The Ethics of Aristotle, 2005, Project Gutenberg, eBook, EBook-No.: 8438, Release Date: July, 2005, Digitised Version of the Original Publication, Produced by Ted Garvin, David Widger, and the DP Team, Edition 10, URL: http://www.gutenberg.org/ebooks/8438 [accessed: 2018-05-10].

[8] Aristotele, Nicomachean Ethics, Volume 1, 2009, Project Gutenberg, eBook, EBook-No.: 28626, Release Date: April 27, 2009, Digitised Version of the Original Publication, Produced by Sophia Canoni, Book provided by Iason Konstantinidis, Translator: Kyriakos Zambas, URL: http://www.gutenberg.org/ebooks/12699 [accessed: 2018-05-10].

[9] T. Gooley, How to Read Nature: Awaken Your Senses to the Outdoors You've Never Noticed. New York, N.Y.: Experiment, 2017, ISBN: 978-1-61519-429-2.

[10] L. W. Anderson and D. R. Krathwohl, Eds., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon, Boston, MA (Pearson Education Group), USA, 2001, ISBN-13: 978-0801319037.

[11] C.-P. Rückemann, "Creation of Objects and Concordances for Knowledge Processing and Advanced Computing," in Proceedings of The Fifth International Conference on Advanced Communications and Computation (INFOCOMP 2015), June 21–26, 2015, Brussels, Belgium. XPS

Press, 2015, pp. 91–98, ISSN: 2308-3484, ISBN-13: 978-1-61208-416-9, URL: http://www.thinkmind.org/index.php?view=article&articleid=infocomp_2015_4_30_60038 [accessed: 2018-05-10].

[12] C.-P. Rückemann, "Advanced Association Processing and Computation Facilities for Geoscientific and Archaeological Knowledge Resources Components," in Proceedings of The Eighth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2016), April 24 – 28, 2016, Venice, Italy. XPS Press, 2016, pages 69–75, ISSN: 2308-393X, ISBN-13: 978-1-61208-469-5, URL: http://www.thinkmind.org/download.php?articleid=geoprocessing_2016_4_20_30144 [accessed: 2018-05-10].

[13] B. Gersbeck-Schierholz and C.-P. Rückemann, "Advanced References: Glue for Value Data," KiM Sky Summit, Knowledge in Motion, September 18, 2016, Sky Summit Meeting, "Unabhängiges Deutsches Institut für Multi-disziplinäre Forschung (DIMF)", Austria, 2016.

[14] C.-P. Rückemann, "Advanced Knowledge Discovery and Computing based on Knowledge Resources, Concordances, and Classification," International Journal On Advances in Intelligent Systems, vol. 9, no. 1&2, 2016, pp. 27–40, ISSN: 1942-2679, URL: http://www.thinkmind.org/download.php?articleid=intsys_v9_n12_2016_3 [accessed: 2018-05-10].

[15] F. Hülsmann and C.-P. Rückemann, "Content and Factor in Practice: Revealing the Content-DNA," KiM Summit, October 26, 2015, Knowledge in Motion, Hannover, Germany, 2015, Project Meeting Report.

[16] F. Hülsmann, C.-P. Rückemann, M. Hofmeister, M. Lorenzen, O. Lau, and M. Tasche, "Application Scenarios for the Content Factor Method in Libraries, Natural Sciences and Archaeology, Statics, Architecture, Risk Coverage, Technology, and Material Sciences," KiM Strategy Summit, March 17, 2016, Knowledge in Motion, Hannover, Germany, 2016.

[17] "Library of Congress Classification Outline," 2018, Library of Congress (LoC) Classification, URL: https://www.loc.gov/catdir/cpso/lcco/ [accessed: 2018-05-10].

[18] M. Saderra Masó, Catalogue of Violent and Destructive Earthquakes in the Philippines, 2006, Project Gutenberg, eBook, EBook-No.: 18556, Release Date: June 11, 2006, Digitised Version of the Original Publication from 1910, URL: http://www.gutenberg.org/ebooks/18556 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/18556/pg18556.txt [accessed: 2018-02-04].

[19] L. Palmieri, The Eruption of Vesuvius in 1872, 2010, Project Gutenberg, eBook, EBook-No.: 33483, Release Date: August 22, 2010, Digitised Version of the Original Publication from 1873, Translator: Mallet, Robert, (1810–1881), URL: http://www.gutenberg.org/ebooks/33483 [accessed: 2018-02-04], URL: http://www.gutenberg.org/files/33483/33483-0.txt [accessed: 2018-02-04].

[20] M. Serao, Sterminator Vesevo (English: Vesuvius the great exterminator), 2014, Project Gutenberg, eBook, EBook-No.: 46658, Release Date: August 22, 2014, Digitised Version of the Original Publication from 1907, Diary of the Eruption of April 1906, URL: http://www.gutenberg.org/ebooks/46658 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/46658/pg46685.txt [accessed: 2018-02-04].

[21] C. Davison, A Study of Recent Earthquakes, 2008, Project Gutenberg, eBook, EBook-No.: 25062, Release Date: April 12, 2008, Digitised Version of the Original Publication from 1905, URL: http://www.gutenberg.org/ebooks/25062 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/25062/pg25062.txt [accessed: 2018-02-04].

[22] W. Hamilton, Observations on Mount Vesuvius, Mount Etna, and Other Volcanos, 2011, Project Gutenberg, eBook, EBook-No.: 35433, Release Date: March 1, 2011, Digitised Version of the Original Publication from 1774, Editor: Cadell, T., (1742–1802), URL: http://www.gutenberg.org/ebooks/35433 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/35433/pg35433.txt [accessed: 2018-02-04].

[23] R. D'Awans, L'Ameublement de l'Hôtel de Pitsembourg au milieu du XVIIe siécle, 2004, Project Gutenberg, eBook, EBook-No.: 11586, Release Date: March 1, 2004, Digitised Version of the Original Publication from 1901, Communication faite en séance du 26 avril 1901, URL: http://www.gutenberg.org/ebooks/11586 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/11586/pg11586.txt [accessed: 2018-02-04].

[24] A. H. C. Gelpke, Ueber die schrecklichen Wirkungen des Aufsturzes eines Kometen auf die Erde und über die vor fünftausend Jahren gehabte Erscheinung dieser Art, 2006, Project Gutenberg, eBook, EBook-No.: 18471, Release Date: May 29, 2006, Digitised Version of the Original Publication from 1835, URL: http://www.gutenberg.org/ebooks/18471 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/18471/pg18471.txt [accessed: 2018-02-04].

[25] "QE: Science: Geology," 2016, Library of Congress (LoC) Classification, URL: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_q.pdf [accessed: 2018-05-10].

[26] "DH: History: General and Eastern Hemisphere: Netherlands, Belgium, Luxemburg," 2016, Library of Congress (LoC) Classification, URL: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_d.pdf [accessed: 2018-05-10].

[27] "QB: Science: Astronomy," 2016, Library of Congress (LoC) Classification, URL: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_q.pdf [accessed: 2018-05-12].

[28] V. V. Vide, Sketches of Aboriginal Life, 2010, Project Gutenberg, eBook, EBook-No.: 33433, Release Date: August 14, 2010, Digitised Version of the Original Publication from 1846, URL: http://www.gutenberg.org/ebooks/33433 [accessed: 2018-02-04], URL: http://www.gutenberg.org/cache/epub/33433/pg33433.txt [accessed: 2018-02-04].

[29] "PS: Language and Literatures: American and Canadian literature," 2016, Library of Congress (LoC) Classification, URL: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_p.pdf [accessed: 2018-05-10].

[30] "UDC 55: Earth Sciences. Geological sciences," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/032958 [accessed: 2018-05-10].

[31] "UDC 9: GEOGRAPHY. BIOGRAPHY. HISTORY," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/068076 [accessed: 2018-05-10].

[32] "UDC 93/94: History," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/068273 [accessed: 2018-05-10].

[33] "UDC 94: General history," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/068284 [accessed: 2018-05-10].

[34] "UDC 52: Astronomy. Astrophysics. Space research. Geodesy," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/027114 [accessed: 2018-05-10].

[35] "UDC 821.111: English literature," 2016, Universal Decimal Classification (UDC), URL: http://udcdata.info/067893 [accessed: 2018-05-10].

[36] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2018-05-10].

[37] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2018-05-10].

[38] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in Proc. INFOCOMP 2012, Oct. 21–26, 2012, Venice, Italy, 2012, pp. 36–41, ISBN: 978-1-61208-226-4, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2012_3_10_10012 [accessed: 2018-05-10].

[39] "UDC Online," 2018, URL: http://www.udc-hub.com/ [accessed: 2018-05-10].

[40] "Universal Decimal Classification Consortium (UDCC)," 2018, URL: http://www.udcc.org [accessed: 2018-05-10].

[41] F. Hülsmann and C.-P. Rückemann, "Summary on Algorithms and Workflows," KiMrise, Knowledge in Motion Winter Meeting, December 12, 2014, Knowledge in Motion, Hannover, Germany, 2014.

[42] "The Perl Programming Language," 2018, URL: https://www.perl.org/ [accessed: 2018-05-10].

[43] "Gnuplot," 2018, URL: https://www.gnuplot.info/ [accessed: 2018-05-10].