Information Integration in a Mining Landscape

GERALD HIEBEL, KLAUS HANKE, GERT GOLDENBERG, CAROLINE O. GRUTSCH, MARKUS STAUDT
University of Innsbruck, Austria

The integration of information sources is a fundamental step to advance research and knowledge about the ancient mining landscape of Schwaz/Brixlegg in the Tyrol / Austria. The approach is applied for the location, identification and interpretation of mining structures within the area. Our goal is to illustrate the use of the CIDOC CRM ontology with extensions in combination with a thesaurus to integrate data on a conceptual level. To implement this integration, we applied semantic web technologies to create a knowledge graph in RDF (Resource Description Framework) that currently represents the available information of seven different sources in a network structure. More sources will eventually be integrated using the same methodology. These include geochemical analysis of artifacts, onomastic research on names related to mining and archaeological information on other mining areas, to research the spread of prehistoric mining activities and technologies.

The RDF network can be queried for research, cultural or emergency response questions, and the results can be displayed using Geoinformation systems. An example of an archaeological research question is the location of mining, settlement and burial sites in the Bronze Age, differentiating between ore extraction, ore processing and smelting activities. For Emergency Services, the names and exact locations of mines are essential in case of an accident within an old mine. Different questions require different subsets of the created knowledge graph. The results of queries to retrieve specific information can be visualized using appropriate tools.

Key words:
Information Integration, Mining Archaeology, Ontology, Semantic Technologies, Geoinformation.

SDH Reference:
Gerald Hiebel, et al. 2017. Information Integration in a Mining Landscape. SDH, 1, 2, 692-699.

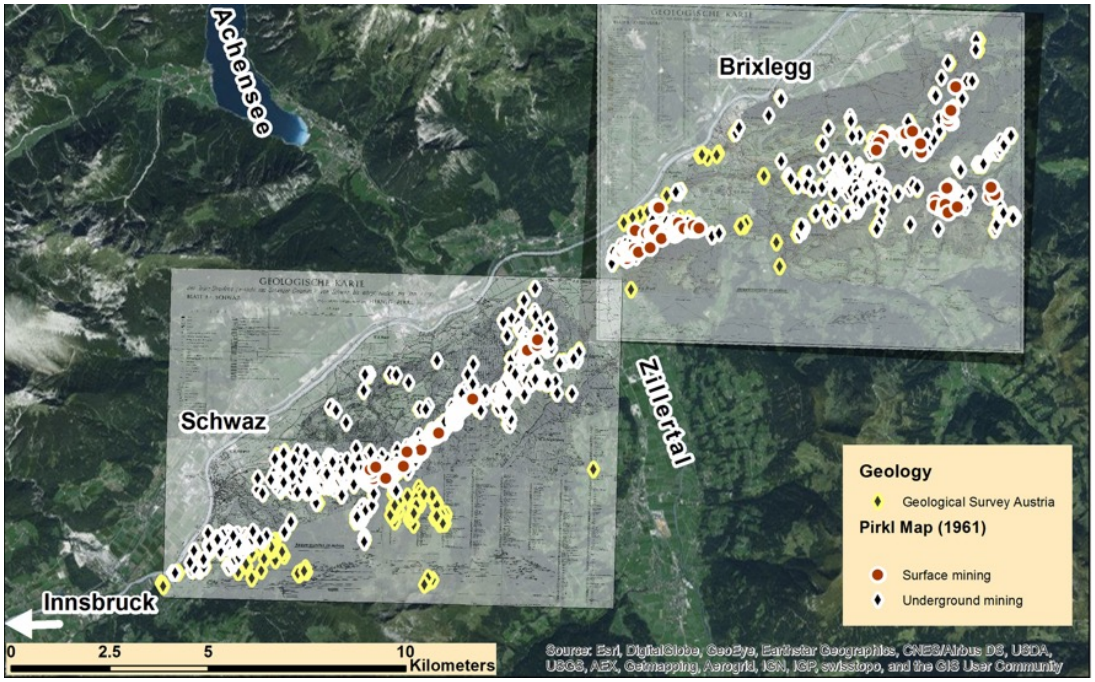
DOI: 10.14434/sdh.v1i2.23231

1. INFORMATION SOURCES

The HiMAT research center of the University of Innsbruck (http://himat.uibk.ac.at) investigates the mining history of the Eastern Alps from prehistory to modern times. Various projects of the research center in the area of Schwaz/Brixlegg target the location, identification and interpretation of mining structures. Geological prospections are a fundamental source of information about structures originating from mining activities. Herwig Pirkl [Pirkl 1961] thoroughly investigated the Schwaz/Brixlegg mining area. The result was a publication describing the geologic and surface

Author's address: Gerald Hiebel, Unit for Surveying and Geoinformation, University of Innsbruck, Technikerstrasse 13, 6020, Innsbruck, Austria; email: gerald.hiebel@uibk.ac.at

structures of the area and containing three geological maps at the scale 1:10000. Two of these maps have been digitized in the course of the work done in the HiMAT research center. Structures identified by Pirkl as underground mining and surface mining have been registered together with their names and coordinates. In addition, information on mining structures provided by the Geological Survey Austria [GBA 2014] has been integrated (figure 1).



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 1. Mining structures identified by Pirkl and the Geological Survey Austria (source: GBA, 2014).

To better locate structures, the high-resolution elevation model of the province of Tyrol was examined for concave and convex surface structures that are in proximity to the structures identified by Pirkl (figure 2). Information about the archaeological sites has been extracted from archaeological literature and from project reports of archaeological prospections and excavations conducted by the HiMAT research center (figure 3). The most recent project, "Prehistoric copper production in the Eastern and Central Alps," contributed significantly to our knowledge of the archaeological sites related to prehistoric mining activities. To document the research in the area, we used the HiMAT database [Hiebel et al. 2013], which records the research activities conducted in the years from 2007 to 2011.

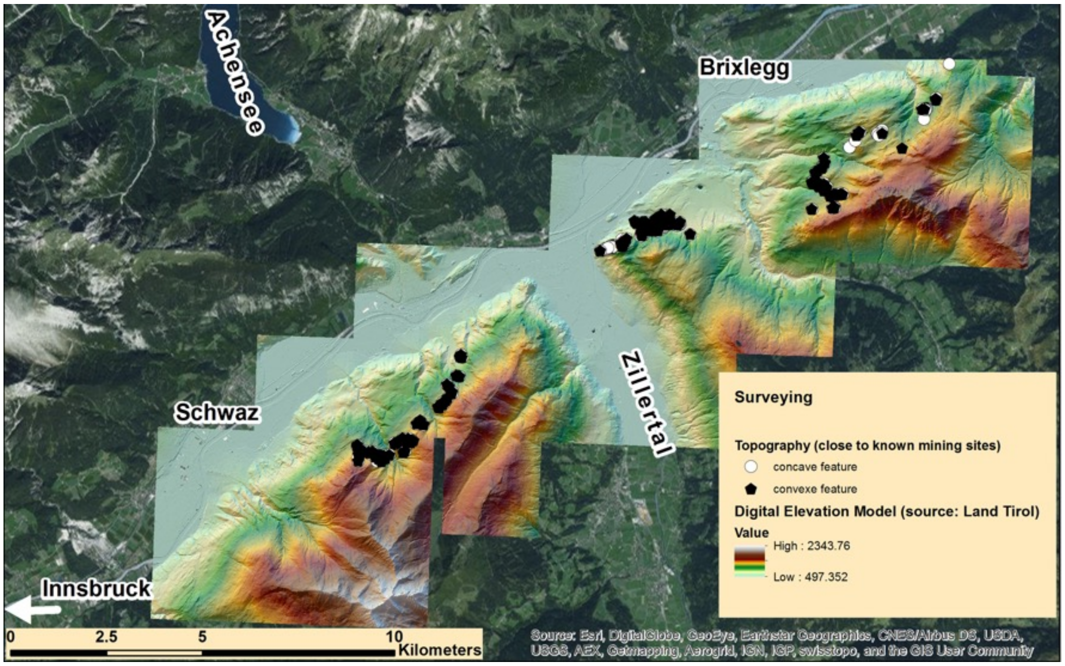

Achensee
Brixlegg
Schwaz
Zillertal
Innsbruck
Surveying
Topography (close to known mining sites)
○ concave feature
⬟ convexe feature
Digital Elevation Model (source: Land Tirol)
Value
High : 2343.76
Low : 497.352
0 2.5 5 10 Kilometers
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 2. Surface structures identified in the high-resolution elevation model of the province of Tyrol (source: Land Tirol 2009).

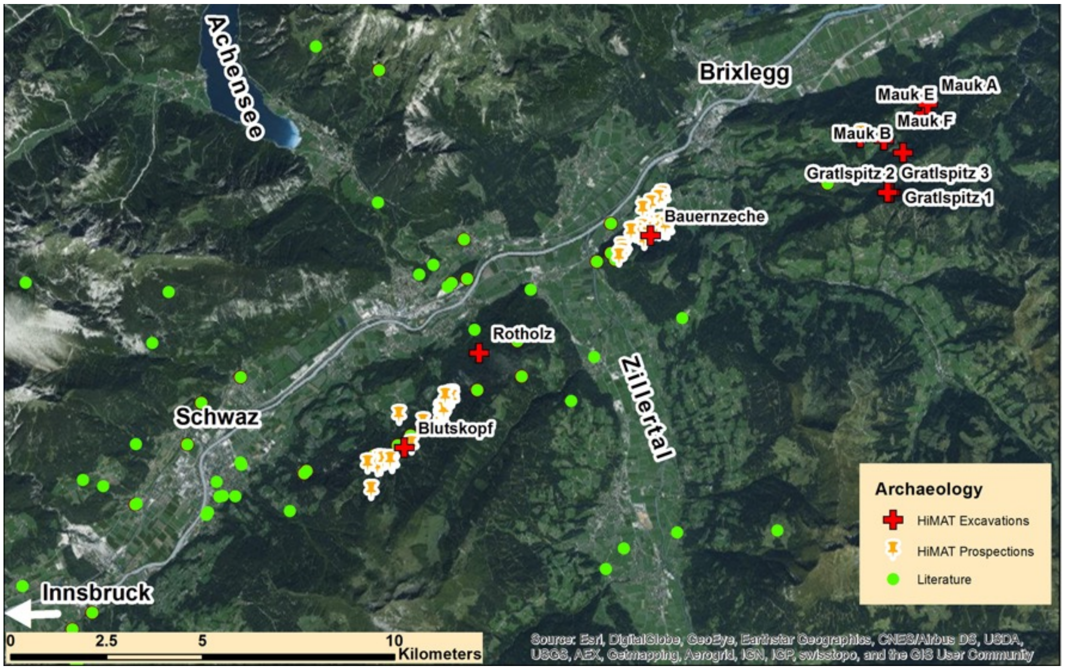


Figure 3. Archaeological sites in the area of Schwaz/Brixlegg (source: HiMAT 2016).

2. INFORMATION INTEGRATION

To integrate the heterogeneous information described in the previous section, we first needed a conceptual model that has the ability to represent the concepts coming from the different HiMAT research domains such as geology, surveying, archaeology, linguistics or metallurgy. The CIDOC CRM ontology [Le Boeuf et al. 2016] was chosen because it is an event-centric data model, and we identified past mining activities and contemporary research activities (which are subclasses of events) as the essential nodes that relate research objects to the documentation and hypothesis created in archaeological research (figure 4). Extensions of the CIDOC CRM [CIDOC CRM 2016] were used to model observations (CRMsci), interpretations (CRMinf), geometric information (CRMgeo) and digital provenance (CRMdig). The classes of the model had to be refined with a thesaurus (figure 5) in order to represent the detailed information in the available documentation and to answer research questions relevant to the domain. The integration of vocabularies originating from different sources has been a serious challenge [Doerr 2006]. Within the DARIAH Infrastructure (http://www.dariah.eu), an approach was developed to integrate terms within a backbone thesaurus and thus create the ability to query upper levels without the need to reach consensus on lower level terms, which is often an almost impossible task to accomplish [Dariah EU 2016].

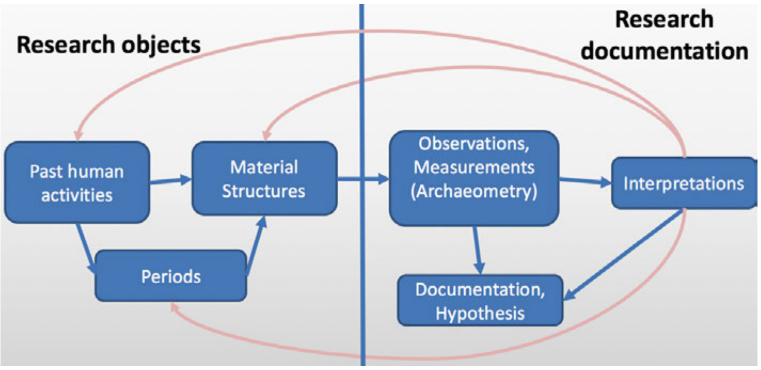


Figure 4. Conceptualizations used for the approach that are represented with CIDOC CRM classes.

We used Karma [ISI 2016], a tool of the semantic web community, to map the information sources to the data model. Figure 6 shows how the original data of the documentation, which must be provided in a structured format (either tabular or hierarchical), was mapped to the formal definitions of the CIDOC CRM ontology. A knowledge graph was created to represent the information, which can be exported in RDF (Resource Description Framework), a data format that is able to relate logical statements within a network [W3C 2014]. RDF is the foundation of the Linked Open Data (LOD) Cloud, where data sets are linked to each other on a global level (http://www.linkedopendata.org). In the continuation of the project, we plan to link the created resources to datasets of the LOD cloud such as Geonames and Wikidata (the human and machine-readable representation of Wikipedia). The thesaurus was created with the Karma tool as well and represented in SKOS (Simple Knowledge Organization System), a data model of the semantic web community for sharing and linking knowledge organization systems, such as thesauri, taxonomies, classification schemes and subject heading systems [W3C 2009].

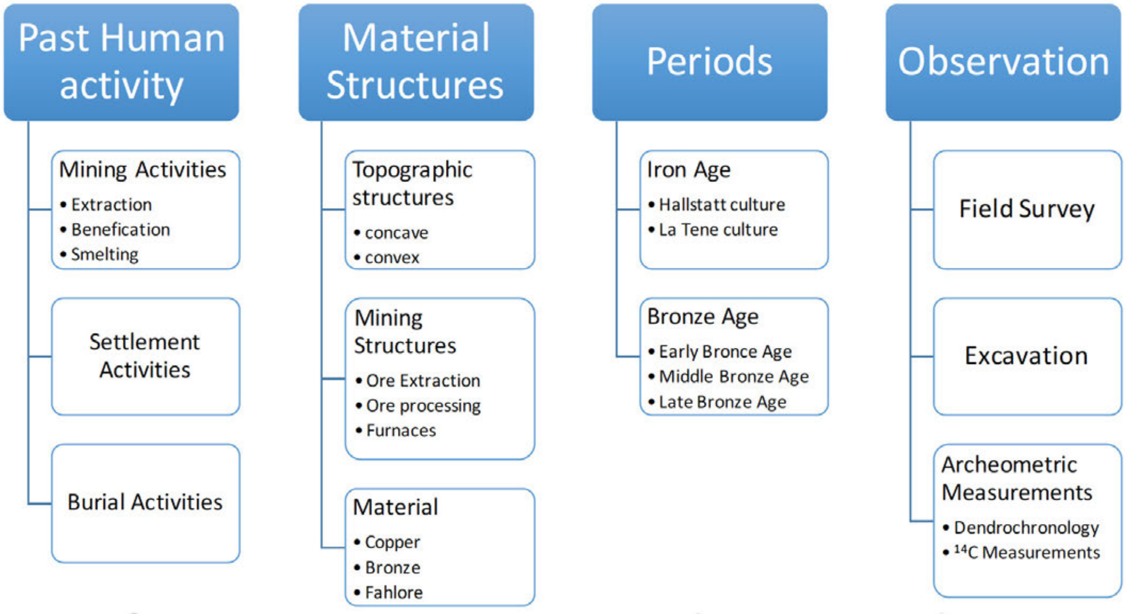
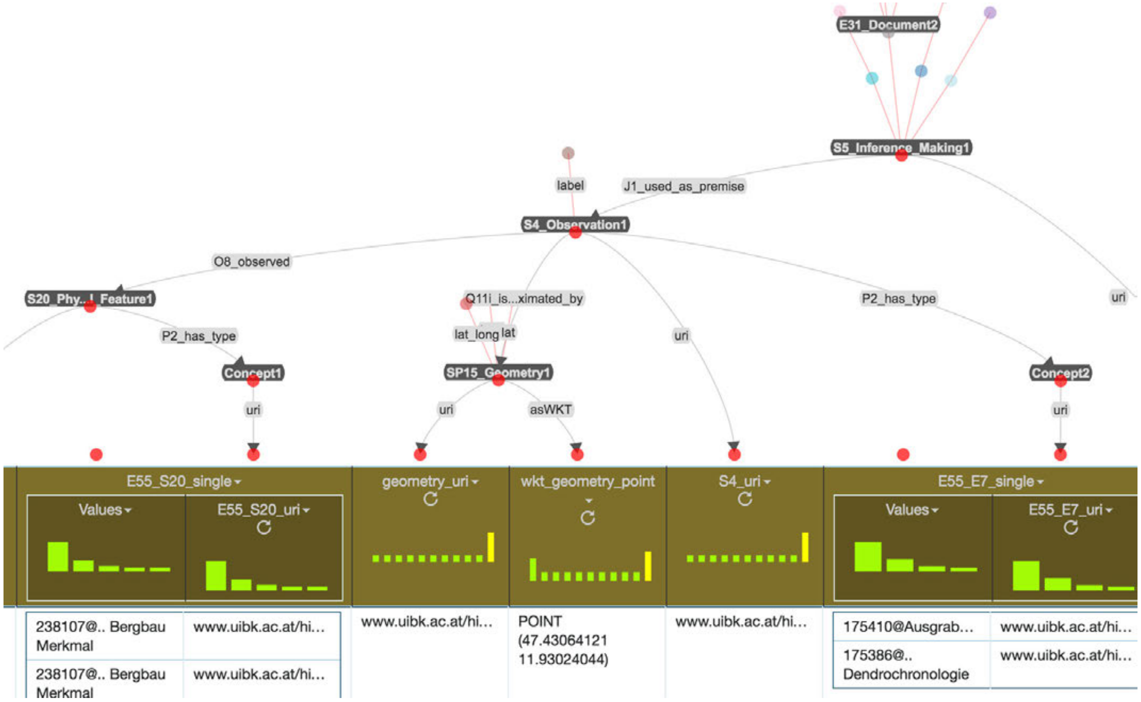


Figure 5. Thesaurus examples to refine the conceptualizations.



Values	E55_S20_uri	geometry_uri	wkt_geometry_point	S4_uri	Values	E55_E7_uri
238107@.. Bergbau Merkmal	www.uibk.ac.at/hi...	www.uibk.ac.at/hi...	POINT (47.43064121 11.93024044)	www.uibk.ac.at/hi...	175410@Ausgrab...	www.uibk.ac.at/hi...
238107@.. Bergbau Merkmal	www.uibk.ac.at/hi...				175386@.. Dendrochronologie	www.uibk.ac.at/hi...

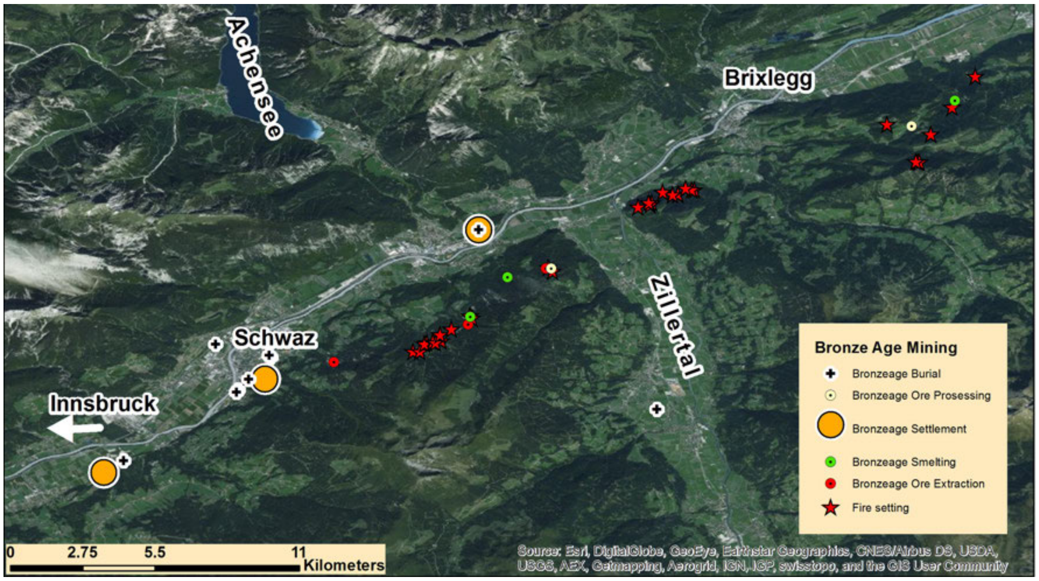
Figure 6. Using Karma to map structured data to the formal definitions of the CIDOC CRM.

After mapping the different information sources and the thesaurus to the common data model, the created RDF structure is placed in a triple store, which is a database to store RDF data. In the triple store, the linking of resources (single information source elements such as a specific underground mine or a concept like the Early Bronze Age) takes place, realizing the actual integration. Resources are either linked on a class level (because they belong to the same CIDOC CRM class, e.g., Observation), on the SKOS concept level (because the same thesaurus term was attributed to them, e.g. , "Early Bronze Age") or on an individual level (because they describe the same material structure object or observation, e.g., "Barbarastollen"). Linking on an individual level is also known as co-reference or entity matching and may involve additional processes to assess the identity of individuals, if no common identifier is available in the different data sources, which is often the case.

3. INFORMATION RETRIEVAL EXAMPLES

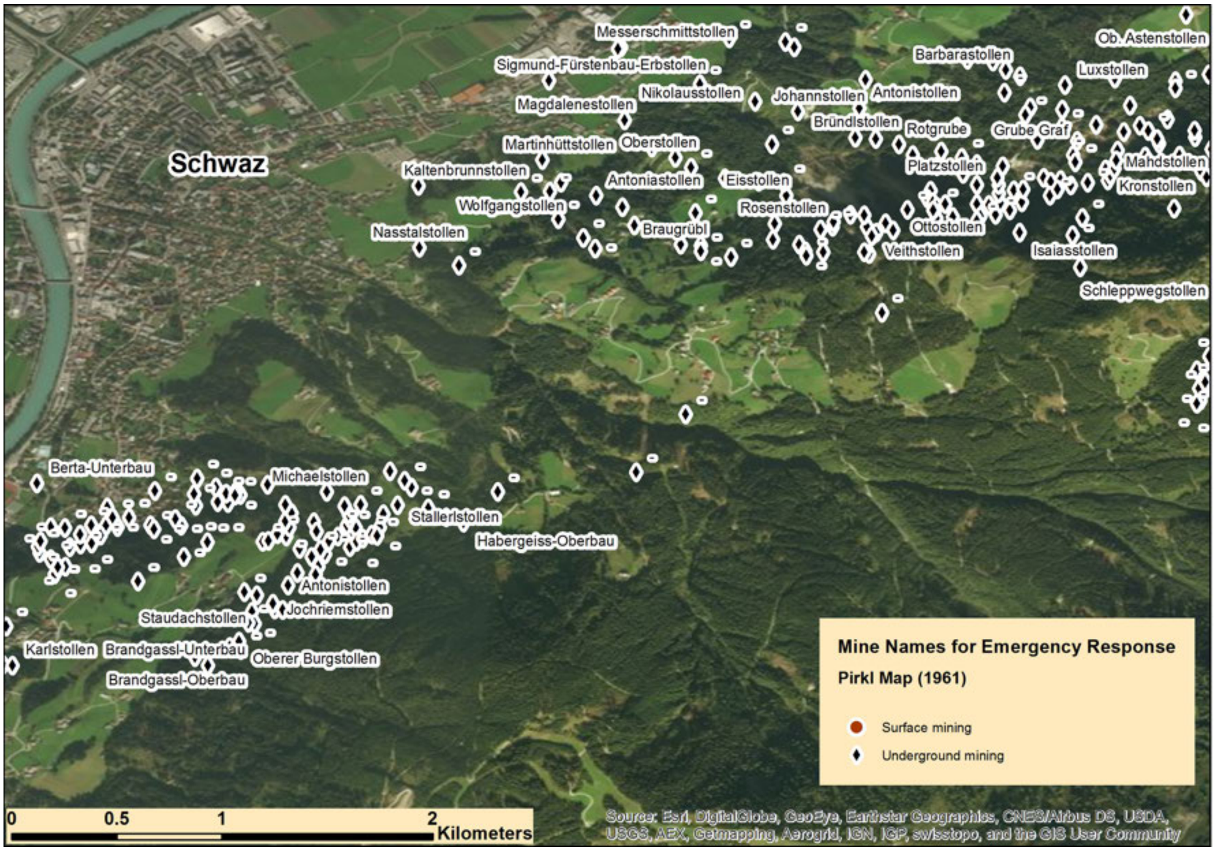
The RDF network of the triple store can be queried using the SPARQL [W3C 2013] query language. To test the integration, we used a model archaeological research question concerning the location of mining, settlement and burial sites in the Bronze Age, differentiating between ore extraction, ore processing and smelting activities. The results of the query were loaded into a Geoinformation system, and a map of known bronze age sites related to mining, settlement and burial activities was created (figure 7).

For Emergency Services, names and exact locations of mines are essential in case of an accident within an old mine. A list of mines containing this information was created from the triple store and given to the Emergency Services. Figure 8 shows the mines and their names on a map. These two application scenarios show how, for different questions, a subset of the created knowledge graph is of interest and that the relevant information can be retrieved and if necessary visualized using appropriate tools.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 7. Map of known bronze age sites related to mining, settlement and burial activities.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 8. Names and locations of mines retrieved for Emergency Services.

4. CONCLUSION AND OUTLOOK

An approach to integrate information available within a mining landscape coming from various sources was developed. A common data model with tools and specifications of the semantic web community was used to perform the actual integration. With specific information retrieval examples, it could be shown that the integration process works and that the triple store can be used to answer specific research questions.

In the current implementation, seven data sources from geology, surveying and archaeology are integrated within the HiMAT database. Further research will apply the methodology to more sources. In the near future, museum exhibition data, a linguistic dissertation about toponyms related to mining and geochemical metal analysis of prehistoric artifacts are targeted. The integrated data will be used to identify material structures or places in LOD sources like Geonames or Wikidata through string matching processes and semantic queries.

5. ACKNOWLEDGEMENTS

The research presented here was financed by the Austrian Science Fund in the course of an Erwin Schrödinger scholarship (J 3646-N15) and by the University of Innsbruck in the course of a graduate scholarship (240346).

6. REFERENCES

Dariah EU. 2016. DARIAH Backbone Thesaurus (BBT) - Definition of a model for sustainable interoperable thesauri maintenance, Produced by the Thesaurus Maintenance Working Group, VCC3, DARIAH EU, http://83.212.168.219/DariahCrete/sites/default/files/dariah_bbt_v_1.2_draft_v4.pdf (4.1.2017)

M. Doerr. 2006. Semantic Problems of Thesaurus Mapping. Journal Of Digital Information, 1(8). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/31/32 (3.11.2016)

GBA. 2014. Digitale Datensätze des Bergbau/Haldenkatasters betreffend ausgewählter Bergbaugebiete im Raum Schwaz-Brixlegg und Kitzbühel-Jochberg, Fachabteilung Rohstoffgeologie der Geologischen Bundesanstalt

G. Hiebel, K. Hanke, and I. Hayek. 2013. Methodology for CIDOC CRM Based Data Integration with Spatial Data. In: F. Contreras, M. Farjas, and F. J. Melero, editors, CAA 2010: Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010 - BAR International. British Archaeological Reports, UK, 547-554.

ISI. 2016. Karma: A Data Integration Tool, http://www.isi.edu/integration/karma/ (3.11.2016)

CIDOC CRM. 2016. CIDOC CRM Compatible models & Collaborations, http://www.cidoc-crm.org/collaborations (9.1.2017)

P. Le Boeuf, M. Doerr, C. E. Ore, and S. Stead. 2016. Definition of the CIDOC Conceptual Reference Model, http://www.cidoc-crm.org/official_release_cidoc.html (6.4.2016).

H. Pirkl. 1961. Geologie des Trias-Streifens und des Schwazer Dolomits südlich des Inn zwischen Schwaz und Wörgl (Tirol), Jahrbuch Geol. B. A. (1961), Bd. 104. 1. Heft, (Wien 1961)

W3C. 2009. SKOS Simple Knowledge Organization System Reference. https://www.w3.org/TR/2009/REC-skos-reference-20090818/ (19.6.2016)

W3C. 2013. SPARQL 1.1 Overview https://www.w3.org/TR/sparql11-overview/ (9.1.2017)

W3C. 2014. Resource Description Framework (RDF) http://www.w3.org/RDF/ (19.6.2016)