International Journal of Geo-Information

MDPI

Article

Taxonomy-Oriented Domain Analysis of GIS: A Case Study for Paleontological Software Systems

Agustina Buccella [1,*], Alejandra Cechich [1], Juan Porfiri [2] and Domenica Diniz Dos Santos [2]

[1] Facultad de Informática, Universidad Nacional del Comahue, Buenos Aires 1400, 8300 Neuquen, Argentina; alejandra.cechich@fi.uncoma.edu.ar

[2] Museo de Ciencias Naturales, Universidad Nacional del Comahue, Buenos Aires 1400, 8300 Neuquen, Argentina; jporfiri@gmail.com (J.P); domenicasantos@gmail.com (D.D.D.S.)

* Correspondence: agustina.buccella@fi.uncoma.edu.ar; Tel.: +54-299-4490-649

Received: 23 April 2019; Accepted: 28 May 2019; Published: 11 June 2019

Abstract: Documenting the paleontological process includes data produced by different techniques and protocols, which are used by paleontologists to prospect and eventually find a new fossil. Nowadays, together with the aforementioned data, a great amount of information is also available in terms of georeferenced systems, including contextual as well as descriptive information. However, the use of this information into a model capable of recognizing similarities and differences is still an open issue within the Natural Heritage community. From the software engineering field, software product lines are models that focus on reusing common assets, in such a way that new software developments are only concern on differentiation relying on already modeled (and implemented) systems. This synergy leads us to apply our taxonomy-oriented domain analysis for Software Product Line (SPL) development, when building systems for documenting the paleontological process. In this paper, we introduce the approach for building such software systems, and illustrate its use through a case study in North Patagonia. Findings show promissory results in terms of reuse.

Keywords: domain analysis; taxonomy; software product line; paleontological domain

1. Introduction

In the geographical area, software applications share behavior and similar information that at first glance would seem impossible. For example, what can be similar between analyzing the geographic distribution of schools of fish in the ocean and identifying sites where paleontological findings have been made? It is evident that both activities share a spatial positioning, a geographic area from which to analyze distributions of specimens, whether they are alive in the ocean or they lived millions of years ago in Patagonia. Spatial positioning leads us to abstract reality as a *set of domains*, all dependent on the geographical domain, which overlap and share certain features.

Domain analysis and reusable software models have come a long way since their first steps in the 1990s. Several efforts have been made in order to provide different mechanisms for improving and maximizing reuse during these last years proposing novel methodologies, techniques and resources. In particular, within domain-oriented methodologies we can cite the Software product line engineering (SPLE) [1–3], which includes two main phases, *domain* and *application*. The first one enables the analysis, modeling, implementing and testing of commonalities and variabilities included in a particular domain. The second one allows the instantiation of the variabilities in order to create particular products. Each product will be the result of all the commonalities and those variabilities chosen by specific requirements of the product being developed. At the same time, each product can contain product-specific services not included in the domain phase.

In this work, we applied the methodology defined in [4,5] as an adaptation of several methodologies widely referenced in academy and industry [2,3,6]. Particularly, the *domain engineering* phase includes the following activities:

- Domain analysis
  Information source analysis (ISA): This activity analyzes sources of information that can support the domain analysis in order to obtain a first set of requirements.

  Subdomain analysis and conceptualization (SAC): The information recovered in the previous process is used to analyze and organize the features or services that the subdomain should offer together with the general features derived from the upper domains. Also, in this process the subdomain must be conceptualized by different software artifacts (such as class models and process models) when it is possible.

  Reusable component analysis (RCA): This process identifies the set of reusable components that could be used to implement the features defined in the last process. It returns a preliminary reference architecture.

- Organizational analysis
  Reuse and boundary analysis (RBA): This activity defines the organizational boundary, commonality, and variability features. Thus, by considering the features specified in the subdomain analysis and conceptualization process and the information from domain experts, the scope of the product line must be defined.

  Platform analysis and design (PAD): This activity builds the reference architecture based on the features defined in the previous activities and processes. The preliminary structure of reusable components defined in the reusable component analysis process is reorganized and refined. Here, each component with its common and variable parts (when necessary) is fully designed.

Although the domain analysis is a common activity when developing software product lines, new challenges remain when we consider larger scale domains as the geography one. This domain involves a wide range of services that can be applied also to other domains such as medicine, e-government, biology, etc. Therefore, when creating a SPL of the geography domain would be impractical, the variability that must be defined to be adapted by all the instantiated products could be really huge. By considering this context, during the last decade, our research has focused on identifying common and variable features by addressing the geographical domain through a bottom-up perspective. We have analyzed the oceanographic domain to focus on the marine ecology subdomain, where we have built a software product line that was reused in different applications and by different users (Centro Nacional Patagónico (CENPAT)—http://www.cenpat.edu.ar/, Instituto de Biología Marina y Pesquera (IBMPAS)—http://ibmpas.org/) [4,5]. We have also addressed the problem of generalizing behavior to later evaluate its application in other subdomains. Currently, through a process that is both top-down and bottom-up, we focus on modeling the paleontology subdomain *with* and *for* reuse: "with" reuse, because we reuse the existing SPL that was modeled for marine ecology, and "for" reuse because we are modeling the aspects of the new subdomain (which potentially extends the original SPL, in addition to reusing it).

In this paper we propose a domain-oriented development based on the use of domain taxonomies that guide the definition and reuse of services. First of all, and as the geographic area has important advances on standardization, we take advantage of standards defined by the ISO/TC 211 committee (Geographic information/Geomatics—https://committee.iso.org/home/tc211) and the Open Geospatial Consortium (OGC—http://www.opengeospatial.org/). These organizations define standards in the field of digital geographic information for building application schemas, representing spatial and temporal information, defining a a geographic service taxonomy, etc. However, as these standards are, in general, abstract and generic, we specialize them into specific subdomains including in this case, marine ecology and paleontology subdomains. At the same time

these subdomains also include standardized information which help to organize the structure of services into the taxonomy. It provides a clear and manageable service structure for facilitating reuse in the construction of these systems. Even, the application of a software product line development process allows us to follow a systematic approach towards maximizing reuse. In this way, the major contributions of this paper are twofold: (1) a service taxonomy development process by standardizing and specializing geographic information as a domain hierarchy, and (2) a software product line development approach, which makes use of the service taxonomy to define reusable functionalities. Our main contribution is that all the generated resources may be applied to develop new product lines in different geographic subdomains.

This paper is organized as follows. The next section describes related work focused on standardization and software development based on service taxonomies. Section 3 presents the background concepts focusing on the geography domain, the paleontology subdomain and existing standards we applied. Section 4 describes how the SPL methodology was applied to the paleontology subdomain, together with the semantic resources used by each activity. Section 5 presents the particularities of the subdomain and describes the process of specializing the taxonomy and defining functionalities. In Section 6 we analyze reuse by applying a graph-edit distance method to obtain preliminary results. Finally, we discuss conclusions and future work.

2. Related Work

Developing software for paleontologists is not a novelty; however, improving developing techniques, i.e., enhancing quality attributes (reusability among others), is a never-ending research. Nowadays, we can find well-established software providers, even those that differentiate their offerings by including geographic characteristics (https://www.gislounge.com/paleontology-and-gis/). However, building new cost-effective software applications still depends on traditional system analysis techniques, which are currently being elaborated to better fit the needs of the paleontology/paleobiology subdomain. Particularly, software for the automatic documentation of archeological and paleontological pieces preserved as collections has received great attention. By pursuing the goal of standardization, many organizations have joined their forces to propose recommendations and even standards. CIDOC-CRM has contributed to build a lingua franca among people who work on natural heritage conservation. For instance, the Ariadne Conservation Documentation System was developed to fulfill the needs for documentation of the department of Conservation of Antiquities & Works of Art (SAET), Technological Educational Institution (TEI) of Athens [7,8]. During this project, researches agreed on the core information. The Ariadne's conceptual design took into account CIDOC-CRM concepts and relations as the minimum of information mandatory for the object's identification. But the Ariadne conservation documentation system was designed and implemented as a relational database; so, dealing with complexity of diagrams potentially hinders the effectivity of the proposal in the real cases. For instance, the concept of *conservation*, as discussed in [7], is very wide. It includes diverse conservation procedures that may change objects depending on the evidence found. Thus, information increases in a large number of entities and relations that can result confusing in some cases.

As another example, Arches [9,10] was created in response to the need for a system that fits the inventory requirements of the heritage field without requiring onerous investment of time and resources by heritage organizations individually. Arches relies on CIDOC-CRM, but also on open data standards; and is designed to access and process geospatial data based on the standards and specifications published by the Open Geospatial Consortium. The information architecture of Arches is organized as Data Management Packages (DMP) that the Arches Server has the ability to implement. DMP are represented as graphs, where each node in a graph corresponds to a CIDOC-CRM Class, and each edge refers to a CIDOC-CRM Property. Spreadsheet templates are provided for mapping concepts, and a visualization tool (Gephi—https://gephi.org/users/download/) supports the manipulation of mappings, while maintaining a machine-readable format.

In the context of the HiMAT project (https://www.uibk.ac.at/himat/index.html.en), CIDOC-CRM has been used for ontological data integration by including spatial data in the integration process [11]. The conceptual part of this process consists of three steps that range from a scope definition over CRM class and properties identification to a thesaurus specification. HiMAT has also used CRMgeo (http://www.cidoc-crm.org/crmgeo/), an extension of CIDOC-CRM that treats space always in combination with time, to model prospection activities, archaeological excavations, and survey and dendrochronological analysis [12].

Currently, several proposals are extending CIDOC-CRM through specific collaborations (http://new.cidoc-crm.org/collaborations). For instance, the Scientific Observation Model (CRMsci—http://www.cidoc-crm.org/crmsci/) is a formal ontology intended to be used as a global schema for integrating metadata about scientific observation, measurements and processed data in descriptive and empirical sciences such as biodiversity, geology, geography, archaeology, cultural heritage conservation and others in research IT environments and research data libraries. CRMarchaeo (http://www.cidoc-crm.org/crmarchaeo/) is an extension of CIDOC CRM created to support the archaeological excavation process and all the various entities and activities related to it. The model takes advantage of the concepts provided by CRMsci, from which it inherits most of the geological and stratigraphic principles that govern archaeological stratigraphy. CRMarchaeo is intended to provide all necessary tools to manage and integrate existing documentation in order to formalize knowledge extracted from observations made by archaeologists, recorded in various ways and adopting different standards. In this sense, its purpose is to facilitate the semantic encoding, exchange, interoperability and access of existing archaeological documentation. Use cases of CIDOC-CRM applications, including Arches, Ariadne and HiMAT, can be found at http://new.cidoc-crm.org/useCasesPage.

At the same time, with respect to the software development itself, to the best of our knowledge, there are no studies applying domain service taxonomies to support the SPL development. However, we can find works defining services within specific domains that could be taken into account here. For instance we can cite the works of *Arizona Dictionary and Taxonomy of Human Services* (https://www.azdes.gov/taxonomy/), the International Foundation for Information Technology (IF4IT—http://www.if4it.com/SYNTHESIZED/FRAMEWORKS/TAXONOMY/service_taxonomy.htm), and [13]. These studies or investigations propose service taxonomies as common languages for improving communication within the software development process.

On the other hand, there is research to develop methodologies for creating taxonomies. For example, in [14] the methodology is defined to develop a taxonomy of mobile applications. Additionally, in [15] the authors propose a process for creating an industry taxonomy within a specific domain that helps in data model development. In the geographic information area, some interesting work defining approaches for categorizing basic operations or features that can be used by any GIS application is described in [16–18]. Another work has been presented by the BEST-GIS project (Best Practice in Software Engineering and methodologies for developing GIS applications) in [16], whose main goal is to define a list of key GIS operations based on the experience of selected users and the contribution of key field experts.

Other efforts to define basic operations in GIS domains have been addressed by organizations for standardization such as the International Organization for Standardization (http://www.iso.org) (ISO) and the Open Geospatial Consortium (OGC). The ISO, by means of the ISO Technical Committee 211, and the OGC has proposed a series of standards to provide rules and methods for creating interoperable geographic systems.

However, even though it is true that there are several efforts describing standard information to define services and assist the development of new products (even for other domains, such as health (ISO 12052:2017 Health informatics—igital imaging and communication in medicine (DICOM) including workflow and data management) or tourism (ISO 21401:2018 Tourism and related services—Sustainability management system for accommodation establishments—Requirements), there is a lack of specific efforts in systematic software reuse. The standards are more related to

interoperability than reuse. In our proposal, the taxonomy-oriented domain analysis includes CIDOC-CRM as one of the standards that guide SPL development—among others such as LIDO, ISO19119, etc. The main goal of the research described here is to make new software development be targeting only differentiation relying on already modeled (and implemented) systems. In this way, we propose to continue the development of a standard definition of services within the geography domain, specializing them into subdomains and focusing on the paleontology subdomain. The objective of the specialization process is to generate a suitable environment within the SPL development by means of defining standard functionalities, services and guides towards a development based on effective software reuse.

3. Background

In this section, we describe the background information needed to understand the bases of our work, involving an overview about the standards of geographic information we are applying (Section 3.1); previous work [5] about a taxonomy defined for the marine ecology subdomain (Section 3.2); and the characteristics of the paleontology subdomain (Section 3.3).

3.1. Standards for GIS: Taxonomy for Geographic Services

Within the geographic information domain we can find several standards covering different aspects about the way the information must be represented, modeled, transferred, digitized, etc. In this work, we start from ISO 19119 std (Services International Standard 19119, ISO/IEC, 2005.) which defines a taxonomy of geographic services. This taxonomy is used as a structure by semantically defining service categories, which help to improve interoperability and reuse. The taxonomy includes five service categories:

- (HI) Human interaction services for management of user interfaces, graphics, and multimedia; and for presentation of compound documents.
- (MMS) Model/information management services for management of the development, manipulation, and storage of metadata, conceptual schemas, and datasets.
- (WTS) Workflow/Task services for support of specific tasks or work-related activities conducted by humans.
- (PS) Processing services for large-scale computations involving substantial amount of data. It contains four subcategories based on the General Feature Model (ISO 19109 std—Rules for Application Schema 19109, ISO/IEC, 2005.):
  - (PS-S) Geographic processing services—spatial
  - (PS-T) Geographic processing services—thematic
  - (PS-Te) Geographic processing services—temporal
  - (PS-M) Geographic processing services—metadata
  - (CS) Communication services for encoding and transfer of data across communications networks.

Here we can see the way these two standards are combined; the ISO 19109 std for defining a 3-layer reference architecture, and the ISO 19119 std for defining the type of services that must be included in each layer. For example, Figure 1 shows a combination that requires the definition of *(HI) human interaction services* in the *human interaction layer*.

Also, in Figure 2a we show a set of service categories belonging to the ISO 19119 std (in italics) together with new services obtained from specific requirements of the GIS systems. For example, we define *(HI-MM) map manipulation* services for grouping those related to interactions with maps, or *(HI-LM) layer manipulation* for services related to interactions with layers. In Figure 2b we show more specializations according to the marine ecology and oceanography subdomains (in gray) for the *(HI-LM4) layer grouping* (connected with a blue arrow) services, and the *(HI-RM3) Hide/show raster manipulation* services (connected with a red arrow).

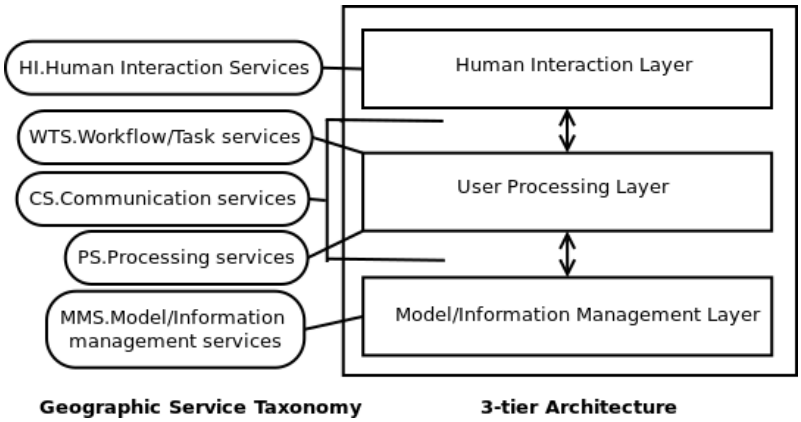


Figure 1. Mapping between ISO 19119 and ISO 19109 stds.

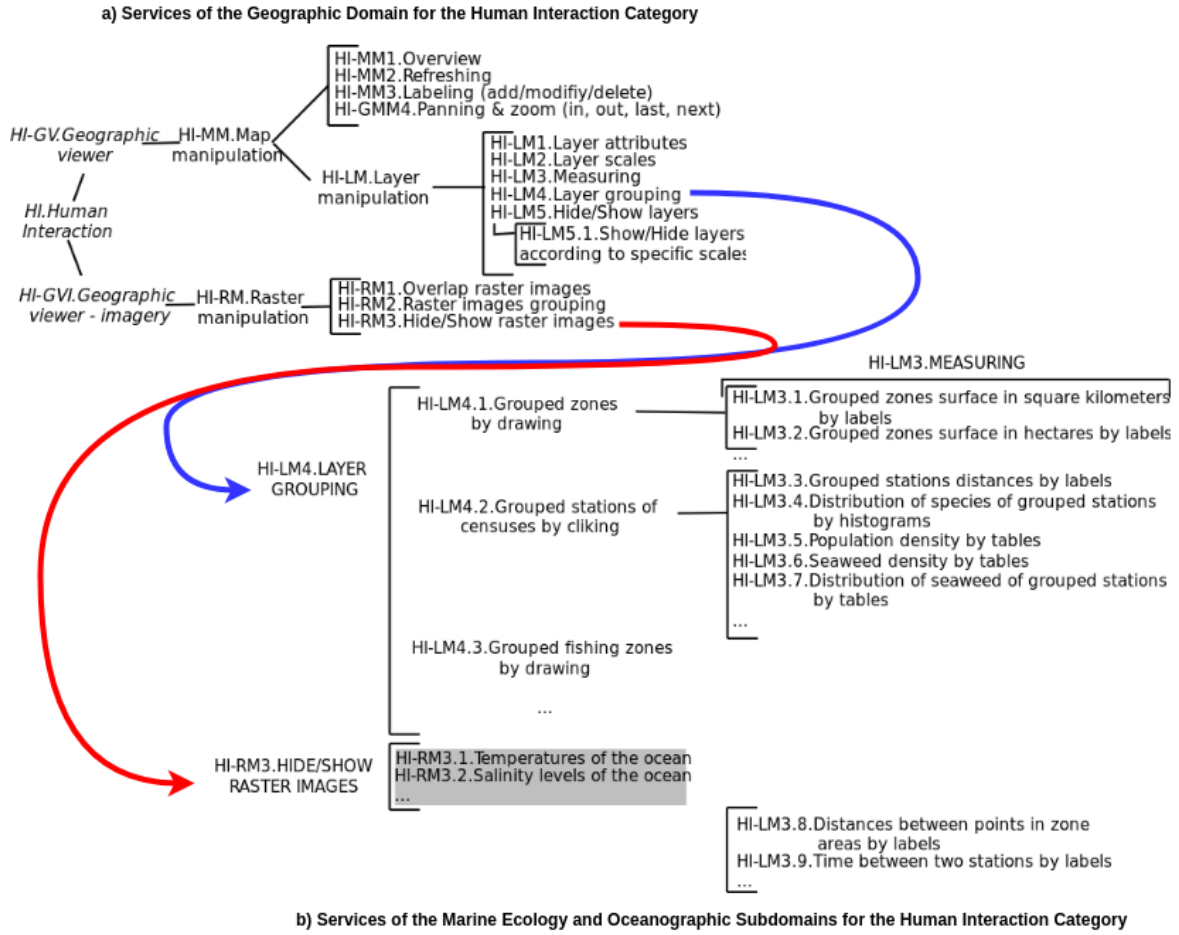


Figure 2. (**a**) Service taxonomy for the geography domain, and (**b**) marine ecology and oceanography subdomains for the human interaction category.

3.2. A Taxonomy for the Marine Ecology Subdomain

In general terms, the marine ecology subdomain involves information about the sea organisms, their particular habitats, their interactions with other sea organisms and external ones, as well as assessing ecosystem impacts from the fishery and pollution. In [5], we have defined a conceptual model and a service taxonomy for supporting software systems in this subdomain; and both elements were used as semantic resources to improve reusability during SPL development. Figure 2b shows an excerpt of the proposed taxonomy, which specializes some services created for showing/hiding layers related to temperatures and salinity levels of specific areas (*(HI-RM3) hide/show rater images*), showing

distribution of seaweed on grouped stations, calculating distances in specific zones or between stations, etc. We refer the reader to [5] for a more detailed description of this taxonomy, as well as its application during SPL development.

3.3. Contextual Information: The Paleontology Subdomain

Paleontology is *a science dealing with the life of past geological periods as known from fossil remains* (https://www.merriam-webster.com/dictionary/paleontology). It is traditionally divided into various subdisciplines, such as micropaleontology (study of generally microscopic fossils); paleobotany (study of fossil plants); palynology (study of pollen and spores); invertebrate paleontology (study of invertebrate animal fossils, such as mollusks, echinoderms, and others); vertebrate paleontology (study of vertebrate fossils, from primitive fishes to mammals); human paleontology (Paleoanthropology); taphonomy (study of the processes of decay, preservation, and the formation of fossils in general), ichnology (study of fossil tracks, trails, and footprints); and paleoecology (study of the ecology and climate of the past, as revealed both by fossils and by other methods). As we can see, paleontology incorporates knowledge from biology, geology, ecology, anthropology, archeology, and even computer science to understand the processes that have led to the origination and eventual destruction of the different types of organisms since life arose. In this sense, the paleontology subdomain overlaps other domains, possibly included in the geographical domain, as we previously discussed.

Similarly to efforts carried out by the OGC, paleontological findings can be characterized for documenting and exchanging purposes according to standard guidelines as well. Particularly, the ISO 21127:2014 (Information and documentation—A reference ontology for the interchange of cultural heritage information) proposes a reference ontology for the interchange of cultural heritage information, which is based on the CIDOC Conceptual Reference Model (CRM) (Conceptual Reference Model Version 6.0—http://www.cidoc-crm.org). It provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. CIDOC-CRM proposes the use of perspectives (What, Where, Who, Events, and When) to analyze key concepts for management and cataloging of pieces, including fossils, as Figure 3 shows. Domain, country and organization specific schemes are suggested for documenting collections, providing metadata in terms of entities (E) and properties (P).

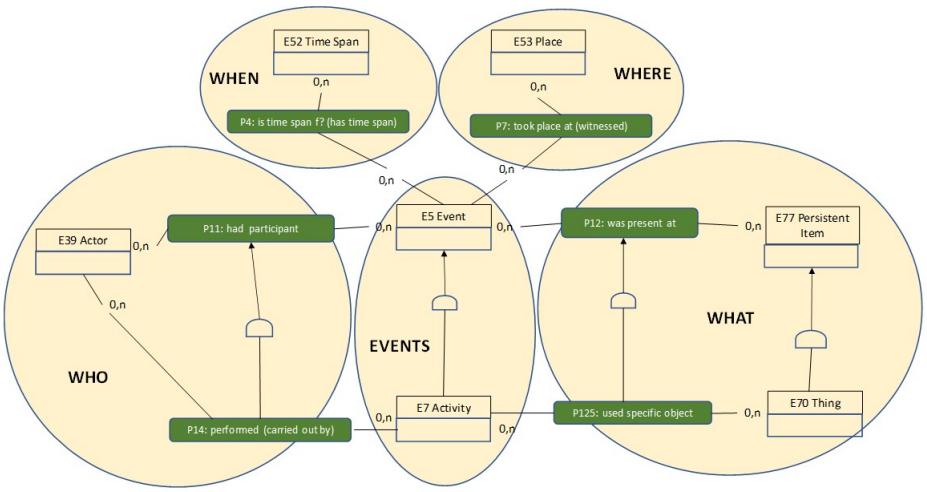


Figure 3. CIDOC-CRM perspectives.

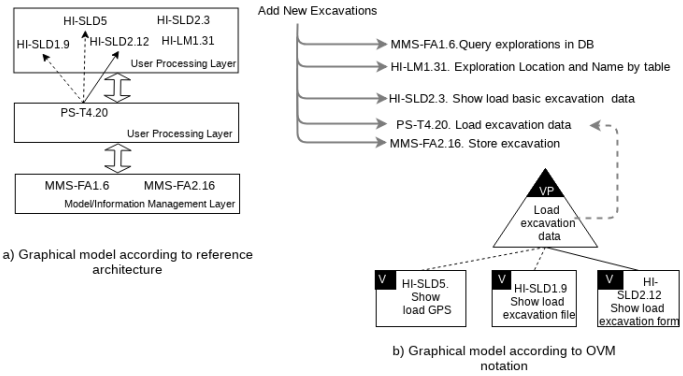
4. Our Domain Analysis Approach in a Nutshell

We describe our approach taking into account the SPL methodology depicted in the introduction section, and the basic elements that differentiate the approach such as the use of taxonomies, as described in Section 3.1. We also consider other semantic resources, such as the conceptual model, the variability model, etc. Figure 4 shows our proposal for domain analysis that begins when domain requirements arrive. Each of the components of the proposal will be described in detail in Section 5. The activities are organized as follows:

- *Information source analysis (ISA)*: During this activity we use the information provided by standards, existing information (in digital and/or paper format) and domain experts. This analysis obtains a first set of requirements (described on Section 5.2).
- *Subdomain analysis and conceptualization (SAC)*: The output of the previous activity is refined in order to generate artifacts (e.g., services) that characterize the subdomain. To do this, two semantic resources are analyzed—the domain taxonomies (Figure 2) and the contextual information. After analyzing requirements information and determining the existence of the relevant domain standards, a domain modeler build the conceptual model (described in Section 5.1) based on domain entities and properties—whether the concepts are standardized from pre-existing standards or the conceptual models are committed to experts. This element is shown in Figure 4 as "Conceptual Model" and was introduced in Section 3.3 for the paleontology subdomain. Similarly, modelers build the domain taxonomy by reusing pre-existing taxonomies (Section 3.1) through refinement, reuse, extension or a combination of these options. For example, what can be similar between (a) finding schools of fish and showing their location on a map and (b) identifying fossil distributions that are being explored in a geographic area? The answer comes from reuse: we can reuse geographic location services in both cases, but display them differently by specializing services corresponding to the human interface; or process information by grouping data differently for each of the cases. This element is shown in Figure 4 as "Taxonomy" and was briefly introduced in Sections 3.1 and 3.2 for the geographic and marine ecology domains. The taxonomy for the paleontology subdomain is further elaborated in Section 5.2.
- *Reusable component analysis (RCA)*: In this activity we build a reference architecture that is used for designing functionality in the way of *functional datasheets*, which are created after analyzing the interactions of selected taxonomy services (referred as "Selected Services" in Figure 4). Each functional datasheet (Table 1) includes five items:

    1. *Id*: an identification,
    2. *Name*: a textual description of the main function,
    3. *Domain*: the domain in which it is included, or it was firstly created,
    4. *Services*: a list of services (from the taxonomy) used to represent the functionality, and
    5. *Variability Models*: a set of graphical artifacts showing the service interactions (as common and variant services). For the last item, any graphical diagram could be used. However, in our work we use a graphical notation based on variability annotation of collaboration diagrams (of UML). The required variability, according to the functionality to be represented, is attached to the diagrams by using the OVM notation [3]. In Table 1 we can see an example of the variability model item for the *Add New Excavations* datasheet in which two graphical models are provided.

Table 1. An example of the functional datasheet for *Add New Excavations*.

Id	FD2
Name	Add New Excavations
Domain	Palentology subdomain
Services	MMS-FA1.6, HI-LM1.31, PS-T4.20, HI-SLD2.3, ...
Variability Models	

Also, each functional datasheet is represented by a set of XML files that allow to automatically analyze these models in search of inconsistencies or incompatibilities when specifying variability (referred as "Variability Model" in Figure 4) [19,20]. Part of the variability model for the paleontology subdomain is discussed in Section 5.3. Then, a *domain component derivation* is performed to create reusable components based on the information defined in the functional datasheets [21] (output arrow labeled as "component structure" in Figure 4). We have defined, in previous research, initial mechanisms that assist in the creation of this component structure [22].

- *Organizational Analysis*: As part of the organization analysis activities, the refined artifacts are used to determine the structure of the software that implements the platform architecture ("Implementation Model" in Figure 4). It models the way in which each refined functional datasheet (based on the taxonomy refinement) is implemented as software components. This platform is then used in the configuration of the products to create the architecture of a specific application. The final output consists of a set of code skeletons ready "for reuse".

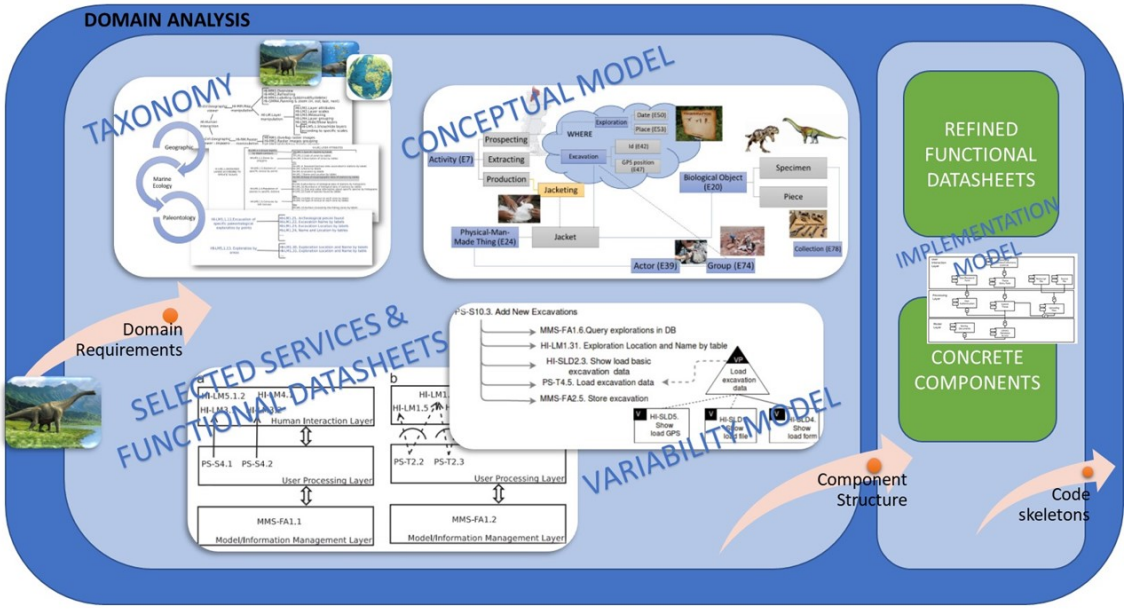


Figure 4. Overview of the taxonomy-oriented domain analysis.

5. A taxonomy to Assist Domain Analysis in SPLs for the Paleontology Subdomain

The service taxonomy for the paleontology domain was developed by following the guides we previously set [4,5], and by an overlapping process that involves conceptual models as well as cataloged services. The following subsections further discuss this process.

5.1. Mapping Conceptual Models

The first task for modeling and processing paleontological information is the definition of the conceptual model. This model organizes the data and their relationships in a consistent structure [23]. The design of this conceptual model involved three information sources: (1) domain requirements, (2) standard information, and (3) information about other domains. In the following subsections, we describe each of them.

5.1.1. Domain Requirements

Our team, along with experts from the Natural Science Museum of National University of Comahue (NSMUnco) (http://extension.uncoma.edu.ar/pagina?id=7), has generated a set of activities that have been abstracted in terms of domain concepts to be applied by the whole paleontology community. The museum is responsible for *prospecting*, *extracting* and *producing* pieces, and therefore, analyzing and storing information about the fossils found in several areas of the Argentinean Patagonia (Neuquen Province). In Figure 5 we show in gray the main entities defined for this domain, and in blue those extracted from the CIDOC-CRM standard. *Prospecting* is a term that paleontologists use that means "hunting for fossils". Each prospecting collects information about the kind of fossils as well as pieces found in the area. In particular, fossils from the cretaceous period, such as the *velocisaurus unicus*, are studied at the NSMUnco. Once they have permission to search in an area that could contain fossils, they hike around looking for fossil fragments on the ground. Eventually, an interesting fossil is located (*Date—E50*). *Extracting* means that fossils are excavated from the ground for study and display in museums. The first step in removing a fossil (*Biological Object—E20*) is to carefully remove the matrix (dirt and rock) that is covering the top of it. Field team members (*Group—E74*) cover the isolated fossil (*Piece*) first with wet paper towels, and then with plaster coated burlap strips. The paper towels protect the fossil from the plaster. The plaster dries into a hard shell or *jacket* (*Physical Man-Made Thing— E24*) that protects the fossil (*Jacketing* as special case of *Production*). The entire *specimen* (or *piece*) is encased in a jacket for its trip back to a lab or museum. Each fossil is given a number (*Id—E42*), so that it can be easily tracked and identified as part of a fossil collection (*Collection—E78*), which corresponds to a particular specimen.

5.1.2. Standard Information

In order to improve data understandability, the model was also defined by mapping the concepts to entities of CIDOC-CRM. As we described in the previous section, these entities are shown as blue rectangles in Figure 5. The paleontological process is abstracted as an *Activity (E7)*, which is specialized into *Prospecting*, *Extracting* and *Production*. Also, the *Specimen* and *Piece* are specializations of *Biological Object (E20)*, and *Jacket* of *Physical Man-Made Thing (E24)*. Other mappings are represented by *Date (E50)* or *Place (E53)* as attributes of the explorations, and *Id (E42)* and *GPS Position (E47)* of the excavations. Then, *Actor (E39)*, *Group (E74)* and *Collection (E78)* are included as in the standard. At the same time, in Figure 5, the *WHERE* perspective of CIDOC-CRM is easily distinguishable from geoinformation stored as attributes of prospecting and extracting, such as *Place (E53)* and *GPS position (E47)*.

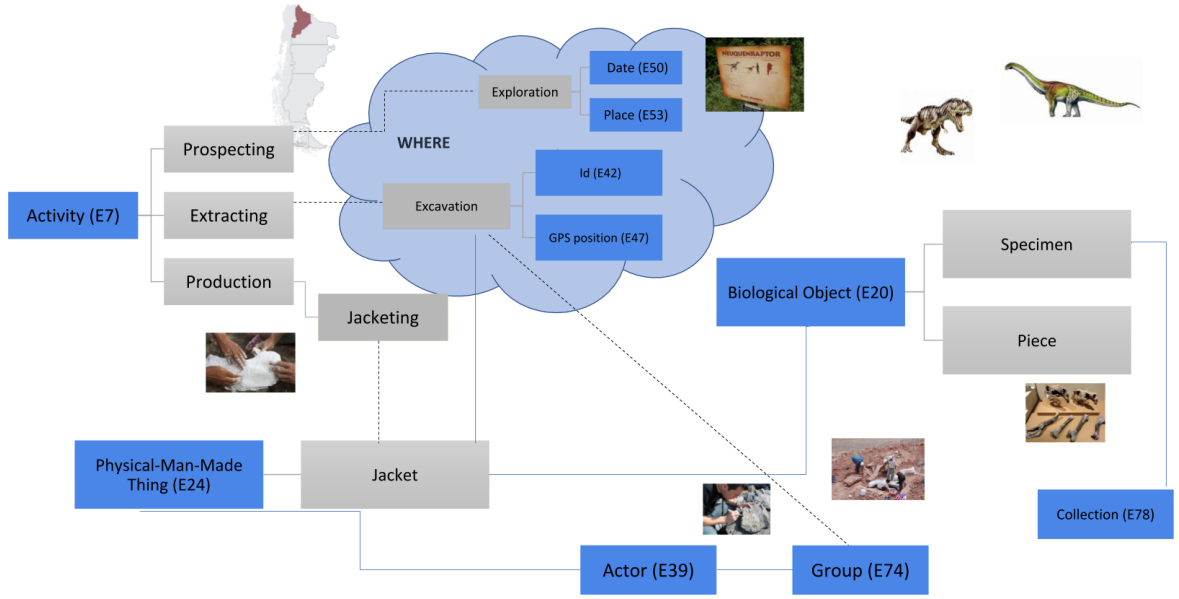


Figure 5. Part of the paleontological conceptual model mapped to entities of CIDOC-CRM.

5.1.3. Information about Other Domains

Finally, we also included information about other domains, such as the geography domain, and thus define some data to be reused [23]. Our first step here was to identify similarities and differences between conceptual elements of the existing domains. In our case, the marine ecology domain had already been analyzed from its geographical and working process perspectives to elaborate a conceptual model. Of course, during this step, it was necessary to collect and analyze information collaboratively with domain experts of the museum, searching for similarities when building this new conceptual model. It was a manual process that took around half a year running meetings iteratively. It also involved some research about specific terms and concepts of the paleontological process, and a generalization process to abstract commonalities.

To illustrate this point, Figure 6 shows part of the previously defined marine ecology's model (as central part of the figure in black) and some of the entities that could be considered similar (indicated by blue arrows). For example, we can see in the marine ecology model that we store information about checklists of species (i.e., censuses). Each census collects information about the population of benthic species living in marine areas. This information is then used for spatial processing in order to obtain the spatial distribution of the species, population variation patterns in different scales, etc. The *Species* class represents a generalization of animals and seaweeds, which are associated to biological or seaweed data respectively, storing specific information about the context and number of species which have been found. All the information is collected from a station (by the *Station* class) that stores a latitude and longitude point in the ocean in which a census was performed. In addition, we defined some classes (in gray in the diagram) that are particular to more general domains, such as the *Bathymetry* class representing measurements of ocean depth by depth contours, and the *Oceanographic_Data* class representing specific information about the ocean conditions. These two classes belong to the oceanographic domain.

As we can see from the figure, geoinformation of the paleontological model can be mapped to data modeled as *Zones* (*Place—E53*) and/or *location* as attribute of *Station* (*GPS position—E47*) in the marine ecology model. Similarly, a *Census* is an activity that involves exploring (very similar to *prospecting*) and clearly maps to the *Activity (E7)* conceptual entity. Finally, *Biological_Data* corresponds to a *Biological Object (E20)*, which relates a *Specimen* to *Animal* (as a specialization of *Species*).

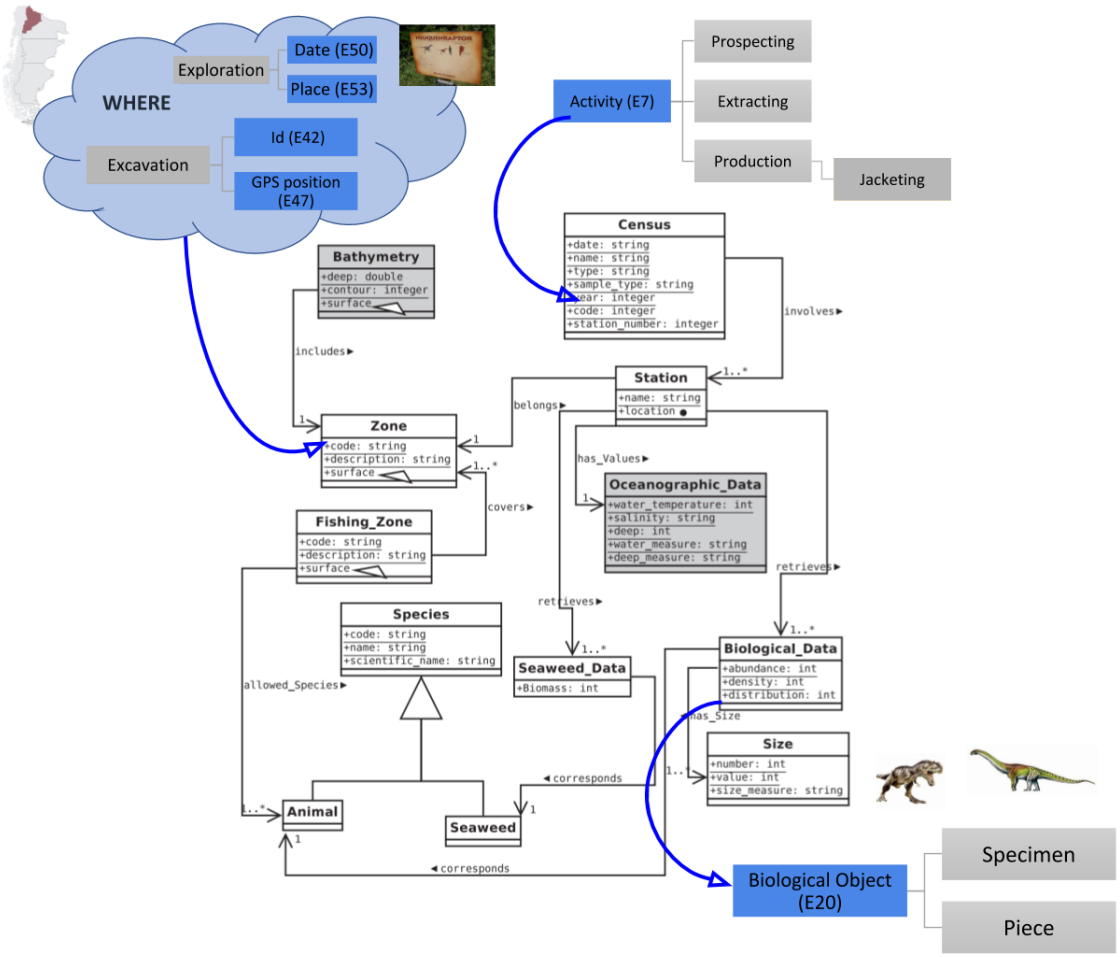


Figure 6. Mapping of the paleontology domain to other conceptual models.

5.2. Creating the Taxonomy in the Paleontology Subdomain

For the creation of the service taxonomy in the paleontology subdomain, we followed the same methodology as presented in [5] by adding new information available and reusing services of the marine ecology taxonomy. This methodology was adapted from the taxonomy development process defined in [15,24] in order to satisfy our requirements. Figure 7 shows the six steps of this process.

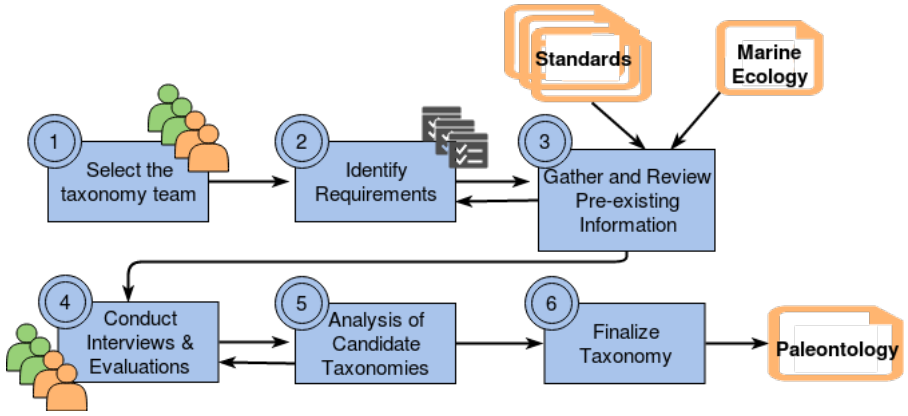


Figure 7. Steps for taxonomy development according to available information.

In the first step, *select the taxonomy team*, we defined the team involving five expert users (paleontologists in general) and three software engineers who participated in the creation of the marine ecology taxonomy.

In the second step, *identify requirements*, the taxonomy team, previously identified, was responsible for specifying the requirements and goals of the new taxonomy. These requirements were obtained from digital and on-paper information related to daily activities of paleontologists in their respective works. As a result, we defined a preliminary list of services necessary for this subdomain. Table 2 groups some of the functionalities or activities performed by paleontologists together with their specific services. For example, jacketing as described in Section 5.1.1, requires specific tools and acid products for helping the extraction without breaks.

Table 2. Some of the preliminary services according to daily activities.

Activities	Services
Recording explorations	assign an exploration code request for the permission for prospecting define the excavation area select the director of the exploration define the dates in which the exploration will be open define the excavation activities ...
Recording excavations	assign an excavation code specify an excavation name register the excavation GPS point attach to an exploration registry define the exploration team define the responsible for exploration record the fossils found ...
Recording jacketing	assign an excavation code assign the preparation type record the acids applied attach to an excavation registry record the biological objects involved record the tools for Excavation label the jacket outside the plastic bag or package ...

This list is subsequently refined and extended during the third step, *gather and review pre-existing information*. This step included the analysis of external information specifically based on standards and general rules defined in the subdomain and upper domains. We have analyzed information from several standards, which are described in Appendix A. It is important to highlight here that standards on cultural heritage information (from the fifth to eighth) are very related to each other. LIDO was defined based on Spectrum and CIDOC-CRM; and the ISO 21127 is the result of the work performed by CIDOC working groups.

Another important external information here is the previous taxonomy in the marine ecology subdomain [5]. This taxonomy was fully analyzed in order to apply the same mechanisms for defining services, reusing similar information and using specific rules for creating the reference architecture (of ISO 19119 std already applied for the SPL in the marine ecology subdomain [4]).

The fourth and fifth steps, *conduct interviews and evaluation* and *analysis of candidate taxonomies*, were iterative processes that defined possible services to be included in the candidate taxonomies. Here, software engineers helped paleontologists describe their tasks by focusing on the way digital information should be managed to support the different processes.

Discussions about paleontologists' concerns with respect to software reuse took several rounds. Current and future roles played by the system were taken into account to relate goals by considering: (1) current situation of computer-assisted processes at the museum, (2) paleontologists' requirements with respect to a given situation, and (3) improvements and potential impact of a

new system. In addition, clear statements about the SPL process development itself were set at the beginning, including perspectives (i.e., aspects such as infrastructure and requirements, social impact by considering the museum a social agent, etc.); scope (the whole processes, some of them, a prioritized list); design features (purpose, existing SPL to be reused, costs, etc.); evolutionary perspectives (different versions of an artifact as it may evolve in time); and abstraction level (generalization/specialization, information hiding, etc.). Domain knowledge evolved from something aware by a paleontologist to something relevant and committed (agreed and adopted possibly changing some process behavior). To do so, each interview had a "knowledge goal"—a knowledge status that the interview helped reach or keep. We used the following strategy to make this process effective [25]:

1. Planning: during this step, we defined the interviews' goals identifying possible sub-goals and an execution order
2. Media and language selection: here, we selected different interactions, such as the use of visual tools, prototypes, similar cases, etc.
3. Cognitive approach: whether we followed an analytic or experimental approach, where cases were run to clarify a point. In some cases, we all together just analyzed material supplied by paleontologists; while other cases required demonstrations on the field (such as jacketing).
4. Social approach: we followed an expert-driven strategy taking into account paleontologists' descriptions during the interviews. These descriptions were committed after an iterative process for agreement. In case cooperation was needed to reach such agreement, we followed a participatory strategy through workshops for knowledge refinement.
5. Conversation techniques: these were selected according to the interviews' goals defined during Planning. For instance, workshops helped us to reach consensus and committed knowledge, as well as validation interviews; meanwhile "brown-paper" sessions were used to identify relevant and aware knowledge. In our case, interviews were progressively structured to incorporate domain knowledge, and identify common and variable features as reusable/required services.

Figure 8 shows an extract of the services defined for the visual part (or user interface) of the functionalities defined in Table 2. These services, associated to specific classes (excavations, explorations and biological objects), define the way data must be shown to the user in order to be created, modified or visualized. For example, excavations may be shown as tables o by labeling the point in which it was performed. Also, biological objects (in this case fossils) may be shown with timelines indicating the dates in which they were found in specific points.

Another type of services are those involving spatial and thematic processing (Figure 9). This division of processing services came from the ISO 19119 in which services involving geographical issues are considered *spatial* or *temporal*, and the others as *thematics*. For example we defined services for calculate surface areas or distances among explorations or excavations; and query or export thematic data.

Finally, as one of the software engineer's tasks, we applied our domain-level approach [4] and defined the domain and subdomains in which we would focus, as shown in Figure 10.

As we can see, we started from the geography domain (with a set of specific standards) and we specialized it into set of subdomains sharing some services. The left part of the figure (Oceanography, Marine Ecology) was performed in previous works as described earlier in the text. In this work, we specialized the cultural heritage subdomain into more specific ones as paleontology and paleobiology. Specifically, our taxonomy ends on the vertebrate paleobiology subdomain. In this way, this domain hierarchy enables us to focus on specific information according to the specification level in which we were working. We can later reuse services of any domain (or subdomain) of the hierarchy to develop SPLs within a manageable set of services.

After this, we redesigned and relocated the services defined previously (in Figures 8 and 9) in order to set them into the correct place of the domain hierarchy. This was done using our *Service Mapper* tool [26,27], which was designed for selecting the best service candidates of the taxonomy according to a required service (or requirement) of the new (paleobiology) subdomain. The selection

process of the best candidates is made by two main activities, *preprocessing* and *indexing and searching*. In the *preprocessing* step, the tool takes a required service (or requirement) as input and uses the WordNet lexical database (https://wordnet.princeton.edu/) to analyze and enrich each of the terms or sets of terms obtained from the requirement and the services of the taxonomy. Then, in the next step, *indexing and searching*, by using the Lucene search engine (https://lucene.apache.org/core/), the documents generated in the previous step are indexed to allow the search engine to find the set of relevant services for the required service. Thus, the software engineers use the Service Mapper tool to obtain the suggested list of candidate services. They must enter the requirement written in natural language, and the tool enriches the text and indexes it to find possible correspondences in the taxonomy. From these candidates, they must decide if the requirement can be fully supported by some candidate, or the required service must be added as a specialization or as a complete new service. The tool is an important support to locate paleobiology services in the correct places of the service taxonomy.

In Figure 11, we can see some of the services of the category *human interaction (HI)* starting from services belonging to the geographical domain (in black and italic), and following with the marine ecology subdomain (in red and italic), and the paleontology subdomain (in blue). Thus, we can observe that services defined in Figure 8 were relocated here by assigning an id according to their place in the hierarchy.

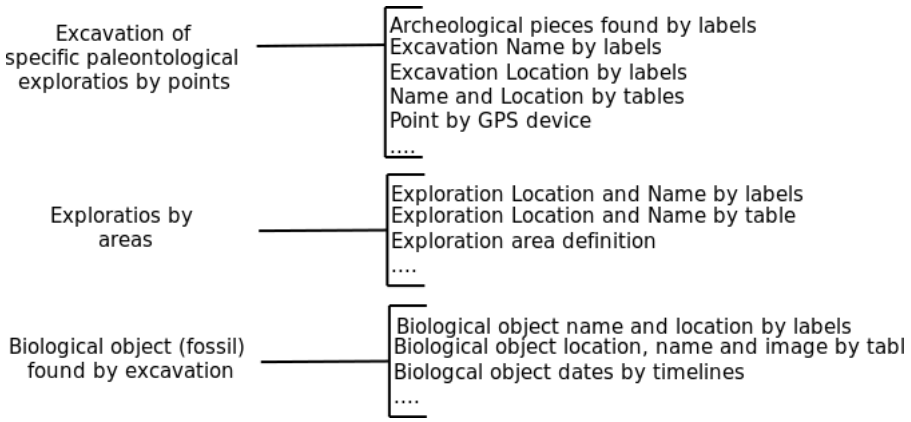


Figure 8. Services defining the way data must be showed to the user.

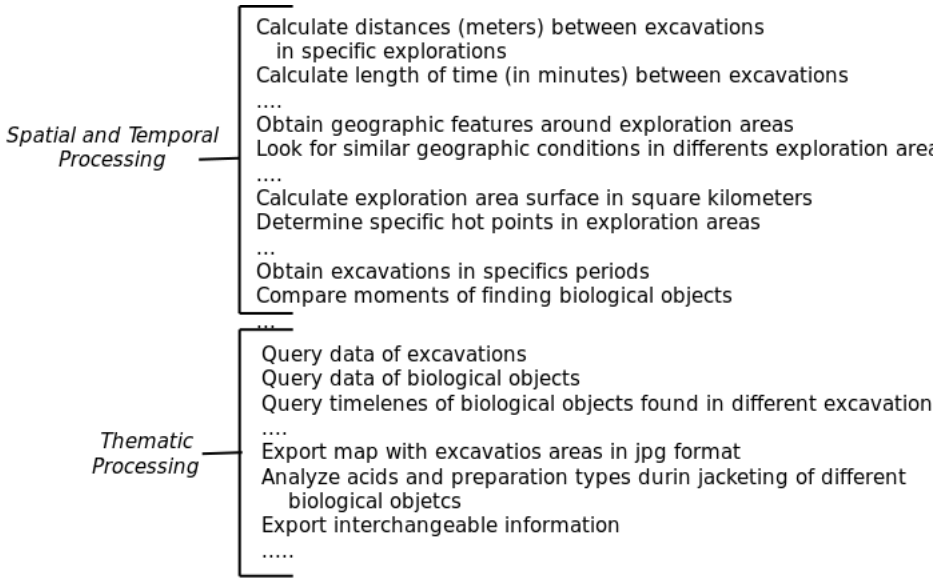


Figure 9. Services representing spatial and thematic processing of paleontological data.

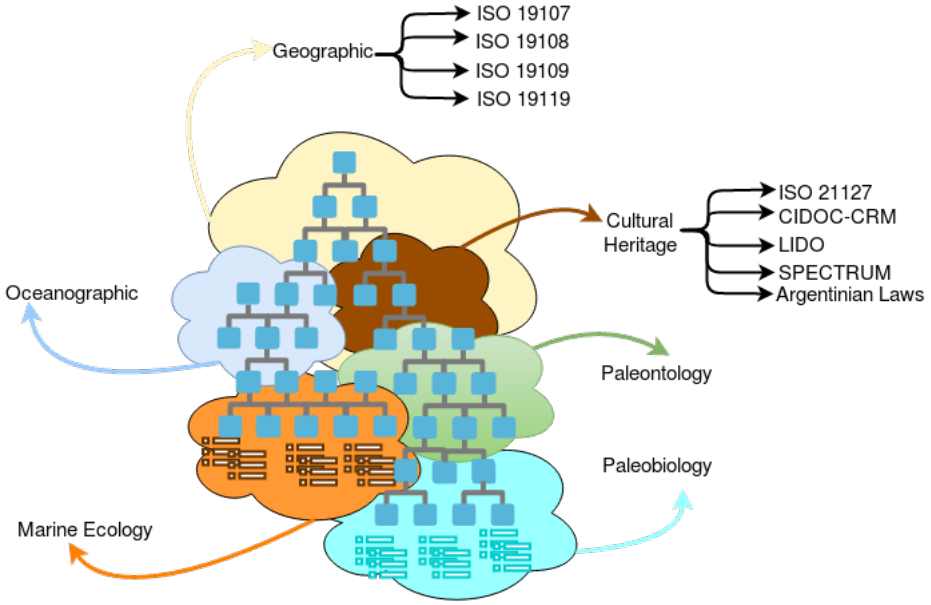


Figure 10. Domains and subdomains included in the analysis.

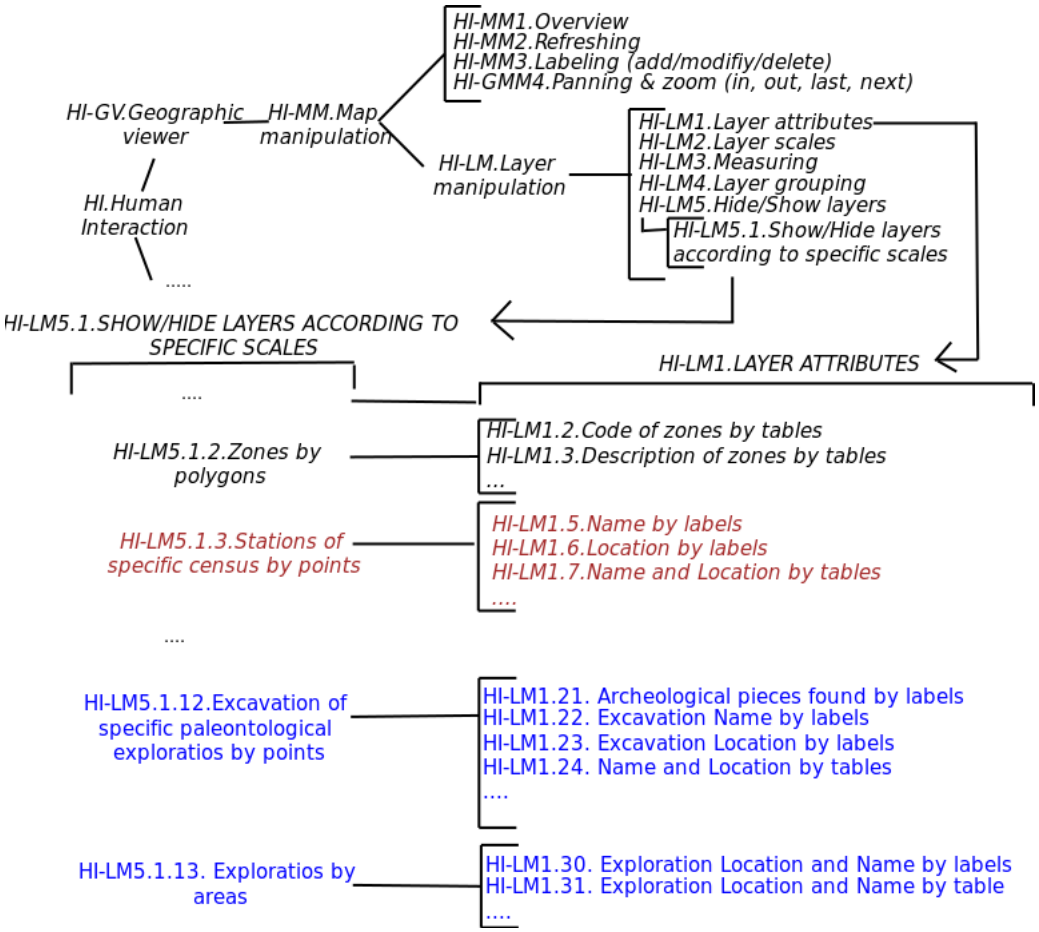


Figure 11. Relocated services in the Human interaction Category.

Another example can be seen in Figure 12, in which the services defined in Figure 9 were relocated as part of the *user processing* category by specializing services of upper domains. Specifically we relocate spatial services as part of the *PS-S.Spatial* service of the ISO 19119 standard.

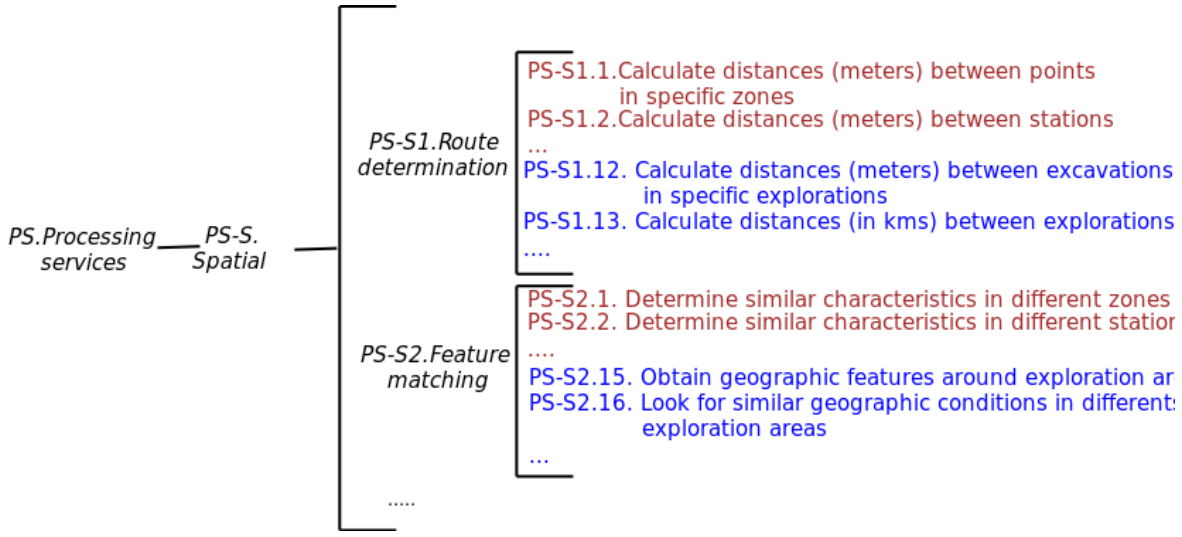


Figure 12. Relocated services in the User Processing Category.

Finally, in the sixth step, *finalize taxonomy*, a review of the taxonomy was performed and final changes were submitted. This step took several rounds of reviews in order to improve and clarify taxonomy descriptions and metadata.

5.3. Guides for Composing Services during SPL Design And Product Implementation

As the service taxonomy in the paleontology subdomain specifies an exhaustive set of services semantically defined, we provide also specific ways in which these services can be combined. These ways are translated as guides defining interactions among the services according to the three-tier reference architecture (Figure 2). These guides must represent generic functionalities that can be used as a basis for implementing a specific function of a particular system. Figure 13 shows the *variability models* item (composed of two graphical diagrams) of the *Add New Excavation* functional datasheet (as presented in Table 1). Figure 13a (Some of the generic functionalities have been simplified in order to use only the services of the service taxonomy described in this article) shows the set of services involved in the *add new excavations* functionality, in which eight services of the service taxonomy are used. This functionality allows to register a new excavation of an exploration already existing in the system. As we can see, on each layer we use the services (In this first figure we only put the identification of the service for clarity and simplicity. The full description of each service are specified in Figure 13b) defined in the taxonomy that provide the semantics for doing the task. Also, we defined variabilities for considering that the load can be done in places without internet connections or GPS devices. Therefore, excavation data (such as geographic coordinates and found objects) is allowed by manually entering information. At the same time, whether is required by any product, the options of loading data from spreadsheets files and/or through devices that contain GPS, are also considered.

Figure 13b shows the same functionality but from a different perspective (as another variability model representation). Thus, the involved services are the same but focused on the order in which they interact with each other. For this functionality, we query the explorations in the DB, select one of them by location and name, show forms for loading basic excavation data, and load data in different formats (GPS, file, or form). Finally, the excavation is stored in the database. The figure also shows optional and mandatory variability included on the way excavation data can be loaded (GPS, file, or form).

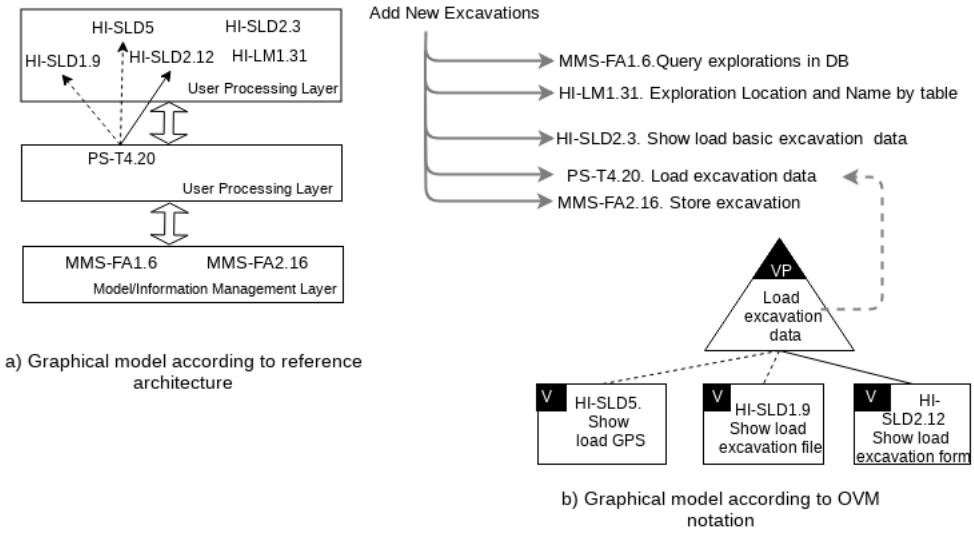


Figure 13. Different diagrams included into the variability model item of the functional datasheet *Add New Excavations*. (**a**) Model according to reference architecture. (**b**) Model according to OVM notation

6. Reuse Analysis

The design of the combined taxonomy by considering existent services and the required ones (Figure 7) allowed us to reuse services completely (as they were defined), partially (with adaptations), or detect the need to redefine a service. After designing, successful reuse will be indicated, in principle, by the existence of a considerable number of reused services, in spite of the complexity of their adaptation.

Traditionally, software reuse community has assessed reusability from different perspectives [28]. One of them addresses functional reuse as behavior already modeled that may satisfy particular requirements. From this perspective, we expect to reuse as many services from the taxonomy as possible. However, the notion of a *considerable* number of reused services may vary depending on several factors, such as the number of outputs a service provides. For instance, reusing 60% of the services may be interesting indeed; however, 60% of fine-grained services providing straightforward outputs, such as the square root calculation, is not comparable to reusing 60% of coarse-grained services, which may provide outputs for supporting an entire business process. In addition, a reused service is usually adapted to a host environment, which can be a traditional software architecture or, as in our case, a SPL platform. Adaptation complexity is assessed in terms of the types of incompatibilities—syntactic, semantic or pragmatic [29], which help select candidate services or components. In our case, and considering that fine- and grain-services are mixed, we took a simplistic view to define *successful reuse* as more than 40% of reused services. This value was subjectively fixed from our previous experience reusing intra-domain services for marine ecology. Adaptation complexity was omitted in our analysis.

In this section, we will analyze reusability by considering the resulting datasheets after finishing the process. Therefore, our analysis will be focused on detecting reusability at datasheet level according to the services that were completely/partially reused.

To do so, and considering that services are actions carried out to accomplish an activity, we assumed that services are tasks to be fulfilled in order to carry out a process. Therefore, having processes in different domains, we have looked at studies in the area of process reusability to analyze the reusability of taxonomy services and/or datasheets. We adapted and extended some similarity functions extracted from studies on similarity in business process models [30–33], which are based on syntactic, semantic, behavioral, and structural measures. In our case, structural measures let us think of reusability in terms of structural changes, and propose a way of reasoning about variability models in datasheets.

To address reusability, we firstly analyzed the reuse of the already existing service taxonomy (provided services) considering the number of services that we have defined for the paleobiology subdomain (required services). Remember that the construction of the taxonomy was the first step in our methodology, so every required service must be already located in some place in the hierarchy (Figure 10). Notice that these are elicited services, which may change during the SPL development as knowledge domain changes; so a *frozen* view is selected as initial required service set.

First of all, we classified services according to the following reuse levels:

- *Completely Reusables Services (CRS)* are those services of the geographical domain or some subdomain, which are reused without modification by the functionalities of the paleobiology SPL. Examples of these services are those that are black and red in Figures 11 and 12, and that are related to this subdomain. These services do not need any type of adaptation to be reused.
- *Reusable Services for Specialization (RSS)* are extended services generated from services already used by other domains. Examples of these services are the blue ones in Figures 11 and 12. In this case, the services need adaptations or extensions to be used by the paleobiology subdomain.
- *New Services for Specialization (NSS)* are new services generated only to this subdomain, and consequently they are not related to services of the other domains. Examples of NSS services are those defined as specializations of the ISO 21127, which were not used by the marine ecology subdomain. Some examples of these services are defined in Table 2, such as *label the jacket outside the plastic bag or package*. In this case, the services need to be designed and implemented from scratch.

Although simplistic, this classification of services is important to draw a first approximation to inter-domain reuse effort. For instance, as CRS services have been already implemented for another subdomain (and they are completely reusable), the reuse effort is practically null. However, when we consider RSS services, the reuse effort might be considerable depending on the adaptation complexity and its implementation. And for NSS services, implementing from scratch might need to consider several rounds of developing tasks.

In this way, this service classification with the aforementioned considerations was used to compare the variability models of datasheets with respect to other subdomains' variability models (In our case each datasheet's variability model was analyzed against variability models of datasheets of the geographic and marine biology domains, which are stored in a repository).

For this comparison, we analyzed the variability model item of the datasheets (Table 1). As we are considering only the variability models represented as Figure 13b, we assumed only one model for each datasheet. For this reason and for simplicity, in the following definitions we refer the variability model simply as "datasheet". Thus, each datasheet was translated by applying the same concepts as presented in [32,33] into directed attributed graphs, according to the following definition:

Definition 1. *(Datasheet) A datasheet is a tuple (S,T,D,A) in which:*
- S is the set of CRS, RSS or NSS services (nodes);
- T represents the variability dependencies (mandatory, optional, alternative, variant, use);
- D represents the variability restrictions (requires, excludes);
- A: $S_1 \cup (T \oplus D) \rightarrow S_2$ *is the set of arcs.* S_1 *and* S_2 *are services, and A represents the relationships among services.*

A datasheet is syntactically correct if and only if it contains at least one service and has strict alternation of the services on each path of arcs from start to end. Thus, by considering the datasheet as a graph, *services* are nodes of the graph, and *arcs* are relationships among services. As an example, in Figure 14 we represent the graph generated from the *Add New Excavations* datasheet shown in Figure 13b.

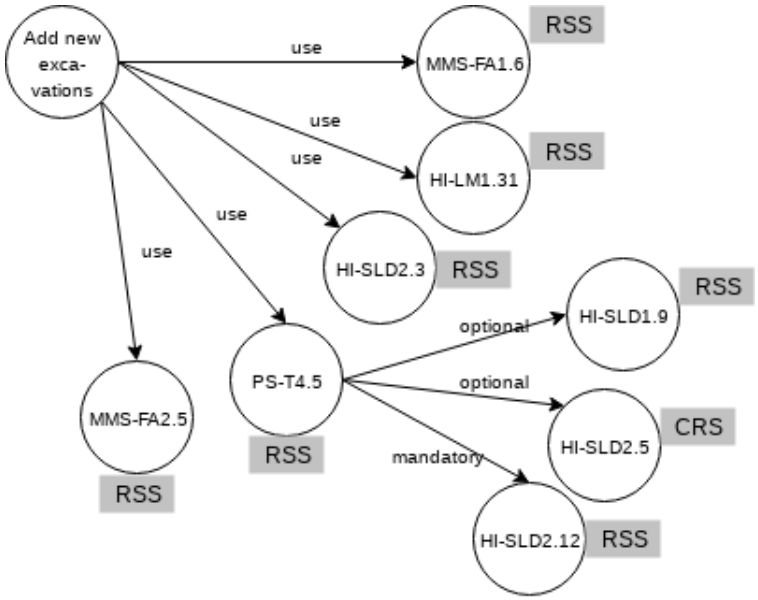


Figure 14. Graph generated from the *Add New Excavations* datasheet, shown in Figure 13b.

For the comparison, we did not apply syntactic or semantic similarity of the text of the services because they are extracted from the taxonomy, so the comparison is straightforward. Therefore, the useful measure to apply here is the structural analysis of the datasheets.

Despite its inherent complexity, measuring the effort (as cost or time) is an important activity. In this section, we only introduce a preliminary analysis for counting the number of services on each category (CRS, RSS, NSS), providing an integrated datasheet level of reuse (DLR); and even our cost model is not developed yet, we will illustrate how DLR can be used for cost calculation.

We assign a similarity score to two datasheets by computing their graph-edit distance [34]. The graph-edit distance between two graphs is the minimal number of graph edit operations that are necessary to derive one graph from the other. We applied this function as defined in [32]. The graph edit operations that we consider are: node deletion, insertion or substitution (a node in a graph is mapped to a node in the other graph with a different service); and edge deletion, insertion or substitution (an edge in a graph is mapped to an edge in the other graph with different variability dependencies or variability restrictions). Formally, we used the graph edit distance defined as (Equation (1)):

$$d(g_1, g_2) = \frac{min}{e_1, ..., e_k \epsilon P(g_1, g_2)} \sum_{i=1}^{k} c(e_i) \quad (1)$$

where $P(g_1, g_2)$ denotes the set of edit paths required to transform g_1 into g_2, and $c(e) \geq 0$ is the cost of each graph edit operation e. In this case, as we are focusing on reuse (and not on costs), we simplified the cost function in order to differentiate only the insertion/deletion/substitution of the three types of services (CRS, RSS, and NSS services). We assigned higher values when RSS or NSS services are involved. In this way, for example, the cost of insertion of a new node (service) is higher when the node involves a RSS or a NSS service.

The use of the graph-edit distance measure, as a mechanism for computing similarities among datasheets of different domains, is aimed to provide extensibility in future validations of the approach. Despite in this work we were focused on the reuse analysis only, the graph-edit distance gives as the possibility to add a cost model in order to differentiate among the edit operations. Even it allows us to include different measure mechanisms according to the variability dependencies and restrictions involved.

In order to carry out our preliminary reuse analysis, we performed three main steps by considering 30 datasheets of the paleobiology subdomain and 60 datasheets from the marine biology subdomain (stored in a repository):

1. We transformed 30 datasheets of the paleobiology subdomain and 60 datasheets from the marine biology subdomain into a graph structure according to Definition 1.
2. We labeled each service of the 30 graphs of the paleobiology subdomain into the three types of services (CRS, RSS, and NSS services).
3. We calculated the cost model of Equation (1) by considering hypothetical costs of deletion/substitution according to the type of service involved: 1 for CRS services, 2 for RSS services, and 3 for NSS services. Although the costs are not real, values intent to show that costs are higher when we have less reuse (In this preliminary analysis we do not consider arcs representing variability dependencies and restrictions). A special case is when we have to insert new nodes, that is, when a graph of the marine ecology domain has nodes that are not present in the graph of the paleobiology domain. These nodes represent services belonging only to the marine ecology subdomain, in this case we assigned the same cost as a NSS service.

In order to clarify these steps, as an example, in Figure 15 we show two graphs corresponding to the two subdomains, paleobiology and marine ecology respectively. These graphs have been built during step 1. Then, in step 2 we labeled each node of the first graph according to the types of services. Finally, the graph edit distance was calculated using Equation (1) in order to find the datasheet of the marine ecology domain, which has the minimum number of edit operations to transform the first model into this one. In this example, we found the *Add new station* functionality as the most similar one. If we apply the equation, according to the costs of the step 3, we obtain a result of 14 corresponding to the substitution of 6 RSS nodes (6 nodes * 2 cost) and 1 deletion of a RSS node (1 node * 2 cost) corresponding to the HI-SLD1.9 node. In the case of the HI-SLD2.5 node we do not have any cost because no edit operation is necessary.

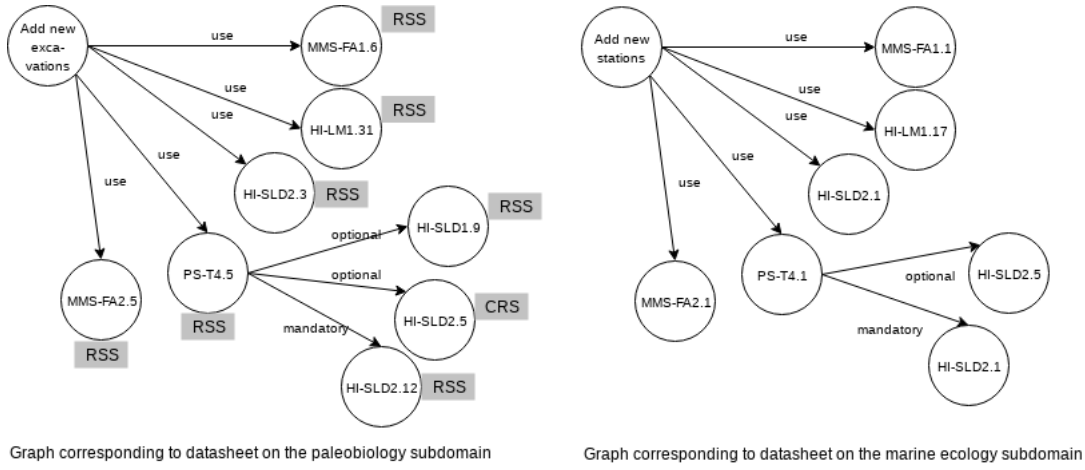


Figure 15. Graphs generated from the datasheets of the two subdomains.

After transforming all the datasheets into graphs (step 1), in step 2 we obtained the number of each type of service associated to each of them. Figure 16 shows graphically these amounts in percentages; where the 31st datasheet in the figure represents the total number of CRS, RSS, and NSS services of the 30 datasheets in order to show the average of service reuse reached. As we can see, more than 80% of the services have been created by specializing other services already defined in the taxonomy; while percentages of *completely reusable* and *new services* is similar (around 17%). Particular cases are represented by datasheets 11 and 19, which only have CRS or NSS services respectively.

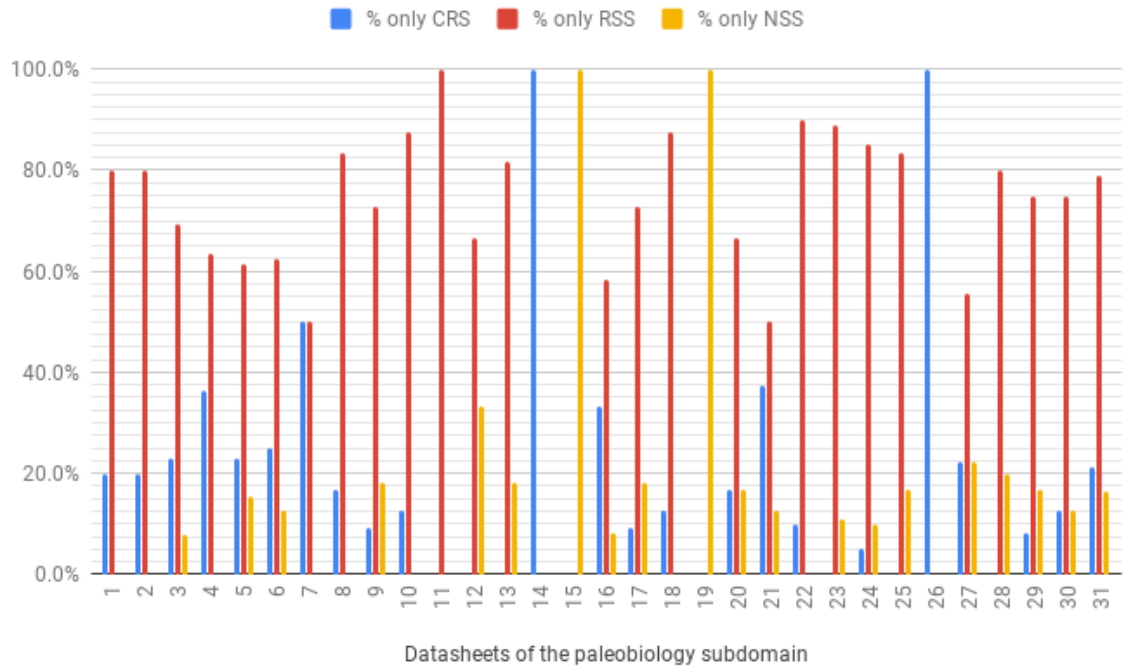


Figure 16. Percentage of the types of services on each datasheet of the paleobiology subdomain.

Following, in step 3 we calculated the Equation (1) for each datasheet of the paleobiology subdomain compared to the 60 datasheets in the marine ecology subdomain. Figure 17 shows the resultant costs by considering the three better candidates, that is, the three first datasheets of the marine ecology domain with the lowest costs when it is compared to each of the paleobiology datasheets. The candidates are denoting the first, second and third datasheets in the repository with better results according to the application of Equation (1). For example, for the datasheet 27 we obtained a cost of 16, which corresponds to the cost of making edit operations to transform the datasheet into one of the datasheets already existing in the repository. The datasheet 27 has 5 RSS services and 2 NSS services with respect to the first candidate datasheet, so according to the third step of our reuse analysis process, we calculate $(5 * 2) + (2 * 3)$ resulting in a cost of 16. Then, a cost of 19 was obtained for the second candidate, and 27 for the last one. The rest of the datasheets (compared to datasheet 27) have a higher result, so they are not considered in the figure.

Particular cases are represented by the datasheets 26 and 14, which have only CRS services (Figure 16), so we found the same datasheet (or functionality) in the marine ecology domain. These are the cases of functionalities related to the geography domain, such as making zoom or calculate distances or areas.

The results obtained in Figure 17 give us final costs that, in spite of being only representative values, provide a panorama of the changes needed for developing the paleobiology's functionalities. The number of edit operations and the associated costs can be then attached to a different cost model in order to measure the effort or time of making the operation possible. At the same time, these costs are comparable to each other and to any other analysis that can be performed when a new SPL is developed for another subdomain.

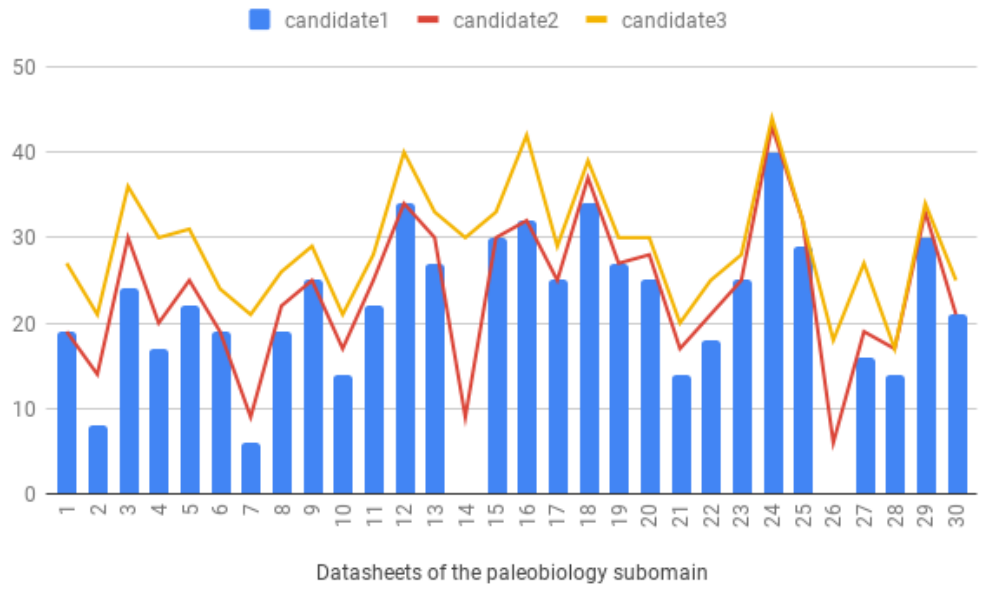


Figure 17. Results of the Equation (1) by applying the three-steps method.

7. Conclusions and Future Work

This paper describes a methodology and a set of guides for software product line development based on service taxonomies. The proposal is mainly focused on the definition of services classified according to a main domain and subdomains involved. Thus, starting from the geographic domain, and considering it as general and complex to be applied, we have performed specializations for adapting services to specific subdomains, in this case the paleobiology subdomain. At the same time, an important aspect is that the services, in addition to be defined by the information provided by expert users, are organized based on the standardized information (about services/activities/functionalities) belonging to each subdomain. In this way, the specialization is aimed at promoting an enabling environment for software product line developments by generating standard operations and guides towards increasing effective software reuse. The study has revealed the following issues:

- The use of previous information available enables software engineers to organize and improve the task of specifying requirements for the paleobiology subdomain: The fact that we had a lot of information available and ready to be used, represented a solid base to define new services and functionalities in the paleobiology subdomain. On one hand, we had a service taxonomy already defined for the geographic and a marine ecology subdomain, which gave us the possibility of organizing the information gathered by domain experts in an already known and applied structure. On the other hand, the use of standards on cultural heritage worked as a controlled vocabulary for all stakeholders providing a common language (which resulted in better communication).
- The definition of services by considering different domains/subdomains was the starting point to improve reuse during SPL development: As we have described throughout and analyzed in Section 6, the service taxonomy and its use in the different subdomains gave us a reasonable and measurable level of reuse of the services on specific functionalities of the product line. The level of reuse identified improved implementing and testing activities during SPL development, because services and functionalities reused were already developed.
- Future software product lines over some other subdomain of the geographic domain can be done by applying the same taxonomy and process: The process of adding new services to the taxonomy belonging to a new subdomain to be included, can be done by applying the same steps proposed in Section 5. These guides, already applied here and in [5], can be also reused as well as all the service taxonomy.

- The reuse analysis gave us an overall view about the real suitability of our development to take advantage of reuse artifacts in a domain: Our analysis provided the general basis to understand the way reuse is reached; and a way to measure it according to the mapping of services (in a taxonomy) and functionalities. However, we must continue working on this analysis by considering real costs of adaptations, extensions and re-implementations of new or reused functionalities.
- New SPL developments might be supported by the same methodology and focused on improving reuse of services and functionalities: The reuse analysis performed in Section 6, in addition to provide a preliminary panorama of the reuse reached, will be useful for building a supporting tool, which helps find the most similar services and functionalities among all lines previously defined. That is what you need when a new SPL is built for another subdomain within the taxonomy.

As future work, we are developing tools for supporting the tasks of managing the taxonomy services and building the datasheets depending on these services. These tools will facilitate the task of working with these resources. At the same time, we are working on extending our reuse analysis, specifically in the graph-edit distance definition, in order to consider variability dependencies and restrictions.

Author Contributions: Agustina Buccella: She has been working on Software Product Line research area for more than 10 years, and this work represents one of the ongoing works in the paleontology domain. Her work here included the writing, the development of the proposal and validation, and the work involved in software development process itself. Alejandra Cechich: She also has been working on Software Product Line research area for more than 10 years, and her main tasks in this article included writting, supervision, and validation of quality, contributions and results. Juan Porfiri and Domenica Diniz Dos Santos: They are biologists specialized on the paleontology domain. In this work they collaborated providing information about the domain, services, and functionalities needed for improving their daily work.

Funding: This research received no external funding.

Acknowledgments: This work is partially supported by UNComa Project 04/F009 of the Universidad Nacional del Comahue.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following acronyms are used in this manuscript:

CIDOC	International Committee for Documentation
CRM	Conceptual Reference Model
CRMsci	Scientific Observation Model
CRS	Completely Reusables Services
CS	Communication services
EOSE	Extended Open Systems Environment
GIS	Geographic Information Systems
NSMUnco	Natural Science Museum at University of Comahue
GPS	Global Positioning System
HI	Human interaction services
ISO	International Organization for Standardization
LIDO	Lightweight Information Describing Object
MMS	Model/information management services
MPL	Multiple Product Lines
NSS	New Services for Specialization
OGC	Open Geospatial Consortium
PS	Processing services
PS-M	Geographic processing services—metadata
PS-S	Geographic processing services—spatial
PS-T	Geographic processing services—thematic
PS-Te	Geographic processing services—temporal

RSS Reusable Services for Specialization
SPL Software Product Line
SPLE Software Product Line Engineering
WTS Workflow/Task services

Appendix A. Standards Related to Paleontology Subdomain

Here we describe the standards applied in this work together with the main information involved:

- ISO 19107 (Spatial Schema International standard 19107, ISO/IEC 2003.): From this standard we analyzed the ways spatial data can be stored together with the set of operations applied to them. For example, each excavation is represented as a GM_Point from this standard.
- ISO 19108 (Temporal schema International standard 19108, ISO/IEC 2002.): Similar to the previous standard but applied to temporal data. Here it is important to store time periods the dinosaurs lived in (TM_Period), specific moments in which fossils had been found (TM_Instant), etc.
- ISO 19109: This standard is the basis of the conceptual schema (Figure 6), which is focused on geographic and thematic classes for representing geographical and non-geographical features respectively.
- ISO 19119: This standard is the core of the taxonomy structure and reference architecture. As we explained in Section 3.1 (Figure 2), we followed a three layer-based architecture by assigning services included within each of these layers.
- ISO 21127 (Information and documentation—A reference ontology for the interchange of cultural heritage information—ISO 21127:2014): We analyzed the reference ontology defined in the cultural heritage information, in which any action about a biological or physical object is represented as an activity (such as Acquisition or Curation). At the same time, this ontology uses geographical information for representing periods, places, etc.
- CIDOC-CRM: As we described in Section 3.3, this standard was analyzed for applying its four perspectives when the services and functionalities were defined. It was applied as a support to the reference ontology analyzed in the ISO 21127.
- LIDO (LIDO—Lightweight Information Describing Objects Version 1.0—http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/): As it is an XML schema for describing museum objects, we analyzed metadata defined here for aligning our information representation.
- SPECTRUM (https://collectionstrust.org.uk/): It is a UK standard for collection management. It specifies how to manage collections and what to do with artifacts during their lifecycle in a collection. Activities defined in CIDOC and ISO 21127 were analyzed with respect to this standard.
- Argentinian Laws: Here provincial and national laws were analyzed in order to adapt processes to the current legislation. For example the Law 25743 (Ley de Protección del Patrimonio Arqueológico y Paleontológico—http://servicios.infoleg.gob.ar/infolegInternet/anexos/85000-89999/86356/norma.htm) defines the way in which the cultural heritage must be acquired, moved, and preserved by specific organizations (such as museums) and the state. These mechanisms support processes defined by international standards by adding specific national information to be considered. As in Argentina the largest number of museums are state-owned, the Argentinian laws are related to specific aspects about how to register the paleontological pieces found, how to ask for exploration and excavation permissions, time ranges in which administrative and legal procedures should be carry out, and penalties in the case of non-compliance. In general, these Argentinian laws do not contradict international standards and some of them refer to some of these standards respect to the way of cataloguing pieces, type of codification, etc. (Decreto 1022/2004—http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do;jsessionid=E7D998DB288023C0F191221C9DA2C53C?id=97432).

References

1. Clements, P.C.; Northrop, L. *Software Product Lines: Practices and Patterns*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 2001.
2. van der Linden, F.; Schmid, K.; Rommes, E. *Software Product Lines in Action: The Best Industrial Practice in Product Line Engineering*; Springer: Berlin/Heidelberg, Germany, 2007.
3. Pohl, K.; Böckle, G.; Linden, F.J.v.d. *Software Product Line Engineering: Foundations, Principles and Techniques*; Springer: Berlin/Heidelberg, Germany, 2005.
4. Buccella, A.; Cechich, A.; Arias, M.; Pol'la, M.; del Socorro Doldan, M.; Morsan, E. Towards systematic software reuse of GIS: Insights from a case study. *Comput. Geosci.* **2013**, *54*, 9–20. [CrossRef]
5. Buccella, A.; Cechich, A.; Pol'la, M.; Arias, M.; Doldan, S.; Morsan, E. Marine Ecology Service Reuse through Taxonomy-Oriented SPL Development. *Comput. Geosci.* **2014**, *73*, 108–121. [CrossRef]
6. Bosch, J. *Design and Use of Software Architectures: Adopting and Evolving a Product-line Approach*; ACM Press Books; Addison-Wesley: Boston, MA, USA, 2000.
7. Naoumidou, N.; Chatzidaki, M.; Alexopoulou, A. "ARIADNE" conservation documentation system: Conceptual design and projection on the CIDOC CRM. framework and limits. In Proceedings of the Annual Conference of CIDOC, Athens, Greece, 15–18 September 2008.
8. Felicetti, A.; Scarselli, T.; Mancinelli, M.L.; Niccolucci, F. Mapping ICCD Archaeological Data to CIDOC-CRM: the RA Schema. In Proceedings of the Workshop Practical Experiences with CIDOC CRM and its Extensions, Co-Located with the 17th International Conference on Theory and Practice of Digital Libraries (CEUR Workshop Proceedings CEUR-WS.org), Valletta, Malta, 26 September 2013; Volume 1117.
9. Carlisle, P.K.; Avramides, I.; Dalgity, A.; Myers, D. *The Arches Heritage Inventory and Management System: A Standards-Based Approach to the Management of Cultural Heritage Information*; Technical Report; World Monuments Fund: Los Angeles, CA, USA, 2014.
10. Myers, D.; Dalgity, A.; Avramides, I. The Arches heritage inventory and management system: A platform for the heritage field. *J. Cult. Herit. Manag. Sustain. Dev.* **2016**, *6*, 213–224. [CrossRef]
11. Hiebel, G.; Hanke, K.; Hayek, I. Methodology for CIDOC CRM based data integration with spatial data. In Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, 6–9 April 2010.
12. Hiebel, G.; Doerr, M.; Hanke, K.; Masur, A. How to Put Archaeological Geometric Data into Context? Representing Mining History Research with CIDOC CRM and Extensions. *Int. J. Herit. Digit. Era* **2014**, *3*, 557–577. [CrossRef]
13. Cohen, S. Ontology and taxonomy of services in a service-oriented architecture. *Archit. J.* **2007**, *11*, 30–35.
14. Nickerson, R.C.; Varshney, U.; Muntermann, J.; Isaac, H. Taxonomy development in information systems: Developing a taxonomy of mobile applications. In Proceedings of the 17th European Conference on Information Systems (ECIS 2009), Verona, Italy, 8–10 June 2009; pp. 1138–1149.
15. Hunink, I.; Rene, E.; Jansen, S.; Brinkkemper, S. Industry taxonomy engineering: The case of the European software ecosystem. In Proceedings of the Fourth European Conference on Software Architecture: Companion Volumem, Copenhagen, Denmark, 23–26 August 2010; ACM: New York, NY, USA, 2010; pp. 111–118.
16. ESPRIT/ESSI Project no 21580. Guidelines for Best Practice in User Interface for GIS, Section 6 List of key GIS Operations. 1998. Available online: https://es.scribd.com/document/169790630/Guideline-for-best-practice (accessed on 31 May 2019).
17. Albrecht, J. Universal GIS Operations for Environmental Modeling. In Proceedings of the Third International Conference/Workshop on Integration GIS and Environmental Modeling, Sante Fe, NM, USA, 21–25 January 1996.
18. Sklar, F.; Constanza, R. The Development of Dynamic Spatial Models for Landscape Ecology: A review and prognosis. In *Quantitative Methods in Landscape Ecology*; Turner, M., Gardner, R., Eds.; Springer: New York, NY, USA, 1991; pp. 239–288.
19. Braun, G.; Pol'la, M.; Cecchi, L.; Buccella, A.; Fillottrani, P.; Cechich, A. A DL semantics for reasoning over OVM-based variability models. In Proceedings of the 30th International Workshop on Description Logics (DL 2017), Montpellier, France, 18–21 July 2017.

20. Pol'la, M.; Buccella, A.; Arias, M.; Cechich, A. SeVaTax: Service taxonomy selection & validation process for SPL development. In Proceedings of the 34th International Conference of the Chilean Computer Science Society (SCCC), Santiago, Chile, 9–13 November 2015; pp. 1–6.
21. Buccella, A.; Pol'la, M.; Cechich, A.; Arias, M. A Variability Representation Approach Based on Domain Service Taxonomies and Their Dependencies. In Proceedings of the 33rd International Conference of the Chilean Computer Science Society (SCCC), Talca, Chile, 8–14 November 2014; pp. 116–119.
22. Arias, M.; Buccella, A.; Cechich, A. Smooth transition from abstract to concrete spl components: A client-server implementation for the geographic domain. In Proceedings of the 1st Symposium of the Argentine Chapter of Geosciences and Remote Sensing Society, Ciudad Autónoma de Buenos Aires, Argentina, 16 June 2016.
23. Pesce, F.; Caballero, S.; Buccella, A.; Cechich, A. Reusing a Geographic Software Product Line Platform: A Case Study in the Paleontological Sub-domain. In *Computer Science–CACIC 2017*; Springer: Berlin, Germany, 2018; pp. 145–154.
24. Choksy, C.E.B. 8 Steps to develop a taxonomy. *Inf. Manag. J.* **2006**, *40*, 30–41.
25. Lankhorst, C.M.L. *Enterprise Architecture at Work: Modelling, Communication, and Analysis*, 1st ed.; Springer: Berlin, Germany, 2005.
26. Arias, M.; DeRenzis, A.; Buccella, A.; Flores, A.; Cechich, A. Classification-based Mining of Reusable Components on Software Product Lines. *IEEE Latin Am. Trans.* **2016**, *14*, 870–876. [CrossRef]
27. Arias, M.; Buccella, A.; Cechich, A. Búsqueda de Funcionalidades basada en Expansión de Consultas para LPS. In Proceedings of the CACIC'16: XXII Congreso Argentino de Ciencias de la Computación, San Luis, Argentina, 3–7 October 2016.
28. Mijač, M.; Stapic, Z. Reusability Metrics of Software Components: Survey. In Proceedings of the 26th Central European Conference on Information and Intelligent Systems, Varaždin, Hrvatska, 23–25 September 2015; pp. 221–231.
29. Gacek, C. *Detecting Architectural Mismatches During System Composition*; Technical Report; University of Southern California: Los Angeles, CA, USA, 1997.
30. Schoknecht, A.; Thaler, T.; Fettke, P.; Oberweis, A.; Laue, R. Similarity of Business Process Models: A State-of-the-Art Analysis. *ACM Comput. Surv.* **2017**, *50*, 52:1–52:33. [CrossRef]
31. Becker, M.; Laue, R. Analysing Differences between Business Process Similarity Measures. In *Business Process Management Workshops*; Daniel, F., Barkaoui, K., Dustdar, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 39–49.
32. Dijkman, R.; Dumas, M.; van Dongen, B.; Käärik, R.; Mendling, J. Similarity of Business Process Models: Metrics and Evaluation. *Inf. Syst.* **2011**, *36*, 498–516. [CrossRef]
33. Dijkman, R.; Dumas, M.; García-Bañuelos, L. Graph Matching Algorithms for Business Process Model Similarity Search. In *Business Process Management*; Dayal, U., Eder, J., Koehler, J., Reijers, H.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 48–63.
34. Bunke, H.; Shearer, K. A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Lett.* **1998**, *19*, 255–259. [CrossRef]