A Methodology for the Design of an Ontology-Based Terminology Resource on an Institutionalised Domain*

Stéphane Carsenty[1]

[1] *French Language Services, Swiss National Bank, Switzerland*

Abstract

This paper describes a methodology used within the framework of the dual dimension of Terminology for the creation of an ontology-based multilingual terminology resource on an institutionalised domain, namely the balance of payments (BOP). Modelling is based on preliminary knowledge, interactions with experts, and corpus analysis. The terminology resource is operationalised and made interoperable, and shall be reusable by translators. In this paper, we go through the different operations performed for modelling and present some challenges faced.

Keywords

multilingual terminology, ontology, ontoterminology, corpus, experts, balance of payments

1. Introduction

Terminology is a discipline studying systematised concepts, which have an expressive side that is most of the time linguistic [1]. It thus possesses a conceptual dimension and a linguistic dimension [2]. From this dual dimension, Terminology acquires its specificity as a scientific discipline [3]. Methodologies and approaches adopted to create a terminology resource should take account of both dimensions.

This paper describes the steps involved in the creation of an ontology-based multilingual terminology resource on an institutionalised domain called the balance of payments (BOP) [4]. That resource shall be human and machine-readable and shall constitute an introduction to the domain for translators and future experts. It is made available in English, in French, and in German. The resource created is an ontoterminology, namely a terminology, whose concept system is an ontology [5]. Knowledge representation encompasses three relations: generic, partitive, and associative relations. The generic relation is the backbone of the ontology. We adopt the approach of the concept as a unit of knowledge created by a unique combination of essential characteristics, after [6]. Each concept is defined intensionally, namely by stating its essential characteristics, or in other words, its generic concept and the characteristics that allow distinguishing it from the latter.

Modelling of terminological information is based on three sources: knowledge acquired through translation and terminology management in our professional activity as a translator and a terminologist in a central bank, interactions with domain experts, and corpus analysis. Basing terminological modelling equally on these three sources makes the originality of our methodology: we are relying neither solely on linguistic analysis, nor exclusively on inputs by experts. A specialised multilingual corpus was built for this research and used both to attest the existence of known terms and for heuristic purposes.

We present hereafter our assumptions (Section 2), before describing the domain of study (Section 3). Section 4 goes through the methodology and its implementation, and Section 5 discusses the results and mentions challenges that were faced, before we conclude (Section 6).

* *4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems", 19-20 June 2025, Thessaloniki, Greece*

Carsenty@gmx.ch

0000-0002-9767-3492

2. Theoretical Framework

As designations used in specialised domains to represent concepts by linguistic means [6], terms are found in specialised texts. In discourse, they can be seen as a point of access to concepts [7]. Specialised texts can be grouped in corpora that are analysed for term extraction.

As for the conceptual dimension of Terminology, a concept is, according to [6], "a unit of knowledge created by a unique combination of characteristics". This unique combination of characteristics is the intension of the concept, namely the "set of characteristics that make up a concept", and an intensional definition is a "definition that conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristic(s)". While the intension of a concept is not explicitly limited to essential characteristics in [6], it is worth noting that all examples given only mention essential characteristics.[1] Moreover, as stated in [8], an intensional definition should "provide the minimum amount of information that forms the basis for conceptualisation". Because we do not think that unessential characteristics belong to this minimum amount of information, we only consider essential characteristics. In our understanding, each concept is thus defined by stating its essential characteristics, or in other words, its generic concept and the characteristics that allow distinguishing it from the latter.

Concerning operationalisation – i.e. computational representation of the concept system [9] –, we represent the concept system as an ontology, i.e. as a "formal, explicit specification of a shared conceptualisation" [10]. The resource is thus a so-called ontoterminology [5].

3. Domain Modelled

We study the domain of the balance of payments (BOP) [4]. The BOP is a branch of official statistics. It encompasses a statistical statement, which supplies information about economic relations between entities linked to a geographic location[2] and the rest of the world. The former are called residents and the latter, non-residents. The BOP belongs to macroeconomic statistics and to international accounts. Macroeconomic statistics are made of aggregates, i.e. groups of objects that can be heterogeneous but possess certain commonalities. These aggregates are recorded in accounts. Heterogeneity is inevitable because economic reality is more complex than the statistical objects built to represent it.

The BOP is an institutionalised domain. Tasks pertaining to its creation (data collection, compilation, presentation and dissemination) are performed by statisticians at central banks or statistics offices. The domain of the BOP is standardised at the international level. Statisticians have to follow special recommendations and accounting principles mainly set out by the International Monetary Fund (IMF). The current reference manuals, BPM6 [4] and its compilation guide [11], were published in 2009 and 2014 respectively. Statisticians will keep using them until 2029. In 2029, they shall implement the principles set out by the new reference manual, namely BPM7, published in March 2025.[3] Countries (e.g. the United States of America [12]) and groups of countries (e.g. the European Union [13]) may use their own terminology. English and French are both used in BOP international reference documents. Nevertheless, diatopic variation exists respectively within the French-speaking and the English-speaking areas.

The BOP is at the intersection of at least three disciplines, namely macroeconomics, statistics, and accounting. Statisticians compile data on macroeconomic phenomena that either occur between entities that are resident in their economy and non-residents (e.g. exports and imports of goods and services, income flows, financial flows), or result from these relationships (financial positions), and they record and present this data in dedicated accounts in the BOP. Furthermore,

[1] "Optical mouse: computer mouse in which movements are detected by light sensors" and "mechanical mouse: computer mouse in which movements are detected by rollers and a ball".
[2] That location can be a country, an economic union or a currency union, and is called "an economy" in the context of the BOP.
[3] See https://www.imf.org/en/News/Articles/2025/03/20/pr25072-imf-and-statistical-community-release-new-global-standards-for-macroeconomic-stats.

concepts that are essential for the BOP are shared with a neighbouring statistic, namely national accounts, and are often described more precisely in the corresponding reference manual [14].

4. Methodology

4.1. Description

We have acquired knowledge of the BOP by translating documents on that topic and by managing a terminology database in the context of our professional activity at the Swiss central bank[4]. Our knowledge is mostly text-based: our acquaintance with BOP concepts and terms is determined – and limited – by the texts we have translated or read in order to understand texts we had to translate. Furthermore, we have widened our knowledge of the BOP by interacting with experts (statisticians responsible for the establishment of the BOP). We are taking part in the text creation process in French, which has a direct influence on the methodology used: our relationship to the BOP is not entirely an external one, we are a kind of "initiate" [15] as we have been in contact with the linguistic means used to talk about the BOP in German, in French, and in English since 2012. We would not be able to produce the BOP terminology based on that experience only, and we still rely on texts to check whether an expression is actually in use, and on specialists to ascertain that it is a term. To sum up, our knowledge of the domain is limited because it is mainly text-based, and we complemented it by consulting experts and with corpus analysis.

For the creation of our terminology resource, we built a specialised corpus on the BOP. Our preliminary knowledge let us make hypotheses about the information to search in the corpus in order to extract term candidates[5]. The work is both corpus-based and corpus-driven: results obtained with first queries led to the elaboration of additional queries and gave insights on aspects and elements, which had not been anticipated.

As for the conceptual dimension, the concept system was modelled as an ontology, based on our knowledge, based on corpus attestations, and based on interactions with experts.

4.2. Implementation

4.2.1. Corpus Design

A specialised corpus was created for this research and organised in five text types that were defined according to the communicative settings they reflect [17], [18] and to the knowledge necessary to understand them (see Figure 1).

[4] www.snb.ch/en.

[5] We use the term "term candidate" instead of "candidate term", which is the preferred term according to [16]. This decision is motivated by the following: the thing that has to be named is a "string of characters that has been collected by means of term extraction but has not yet been selected (…) to be considered for inclusion in a terminological data collection" ([16]). In other words, at the time we are considering it, we are not sure that the string of characters in question is actually a term. We would consequently interpret "candidate term" as an elliptic expression denoting elements of the lexicon that are "candidate for the status of term". Some of them will ultimately not be included in the resource. Based on this interpretation, the head of a noun phrase being its last element in English, we consider it preferable not to use an elliptic expression, and consequently we decided to place "candidate" as head, and thus in last position. At the same time, we would be very interested in a discussion of the reasons that have led to select "candidate term" as the preferred term in [16].

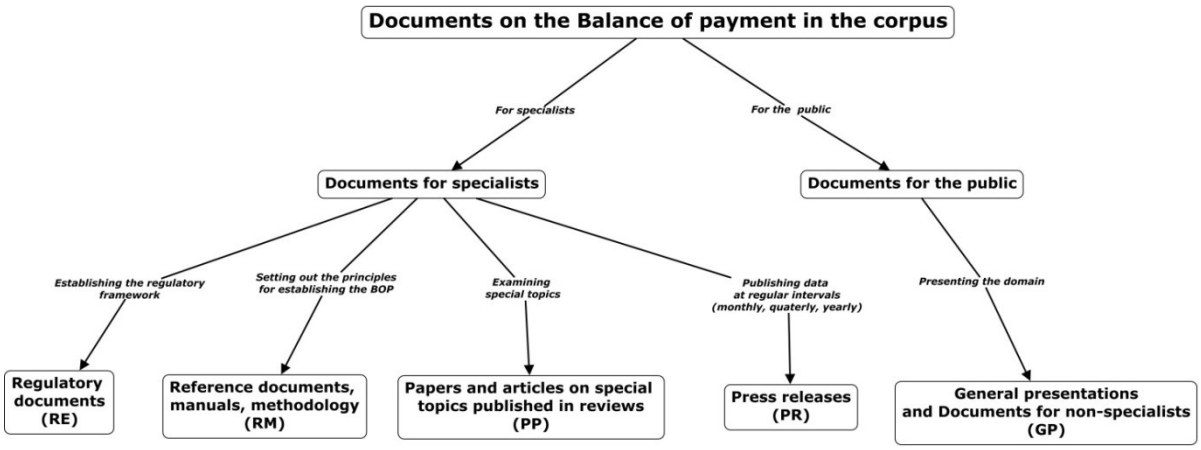

Figure 1: The five types of documents in the BOP corpus

The five types are **Regulatory documents** (**REs**, regulatory issues and laws for establishing the BOP), **Reference documents, manuals, methodology** (**RMs**, principles of the domain, i.e. statistical and methodological topics like data sampling, compiling, computing, and conducting of surveys), **Research papers (PPs)**, **Press releases** (**PRs**, publication of data on the BOP at regular intervals), and **General presentations for non-specialists (GPs)**.

The corpus encompasses 656 documents, with about 29 million characters and 5 million tokens. Texts were published by a central bank, a statistics office or an international organisation like the IMF, in English (48%), in French (38%) or in German (14%), between 2009 and 2024.[6] Central banks, statistics offices, and international organisations publishing documents on the BOP are institutions acknowledged for their expertise [15] in that domain. As these documents have been published, they correspond to authentic communicative situations. In all settings, the authors are experts. There is a predominance of RMs: they account for about 10% of the number of documents but for more than half the size of the corpus. RMs have the aim of standardising the domain.

We collected all texts on the websites of authoring institutions and checked with the latter whether we had missed relevant documents. We can thus assume that the corpus covers the whole range of communicative situations that exist in the specialised domain being studied [19]. Its representativeness is thus qualitative, based on typicality and specialisation [20].[7]

4.2.2. Corpus Workflow

We focussed on RMs in English because these documents correspond to a setting “expert to expert” or “expert to intiate” and aim at standardising the domain, and because English is the language, in which this standardisation takes place.[8]

Morphosyntactic patterns in English for queries in AntConc 4.2.4[9] [21] were defined based on known terms. These patterns are shown in Table 1.

[6] The time frame 2009-2024 is determined by the period of validity of the current BOP reference manual [4].
[7] All data and material can be found in the GitHub folder dedicated to this research: https://github.com/SCarsenty/Ontology-based-terminology-of-the-BOP.
[8] In the BOP corpus, there are 39 RMs in English, 15 in French, and 8 in German.
[9] https://www.laurenceanthony.net/software/antconc/releases/AntConc424/.

Table 1
Morphosyntactic Patterns Used for Queries in AntConc

No.	Term Candidate in English	Pattern to be Searched
1	transaction	_NN
2	institutional unit	_JJ _NN
3	capital account	_NN _NN
4	foreign direct investment	_JJ _JJ _NN
5	international investment position	_JJ _NN _NN
6	insurance technical reserves	_NN _JJ _NN
7	money market fund	_NN _NN _NN
8	balance of payments	_NN _IN _NN
9	balance of international payments	_NN _IN _JJ _NN
10	balance of payments statistics	_NN _IN _NN _NN
11	net incurrence of liabilities	_JJ _NN _IN _NN
12	international merchandise trade statistics	_JJ _NN _NN _NN
13	net acquisition of financial assets	_JJ _NN _IN _JJ _NN
14	other changes in volume account	_JJ _NN _IN _NN _NN

Each query returned a list of term candidates. Successively extending and shrinking the cluster size allowed capturing additional term candidates, and terms were inferred based on the researcher's knowledge.[10]

All term candidates were then submitted to a first selection and validation process based on our knowledge. Firstly, we rejected pleonasms (like **financial liabilities*, all liabilities having the essential characteristic of being "financial" in the context of accounting, as explained in [14]) and usage variants that could be confusing (like **net acquisition of assets* in the context of the financial account, instead of the standard term "net acquisition of financial assets", because only financial assets are relevant in the financial account). As we had focussed our search on RMs, this made us aware of the fact that reference documents and manuals may provide term candidates that we should not select as terms.

Secondly, we cleaned class names, i.e. expressions that do not designate concepts, but classes of things that may be of different natures. As mentioned in Section 3, statisticians group things in aggregates. Designations matching patterns like "_NN and _NN"[11], "_NN not included elsewhere" /

[10] For details on queries and results obtained, see https://github.com/SCarsenty/Ontology-based-terminology-of-the-BOP/tree/main/Corpus/Queries%20on%20English%20corpus.

[11] This pattern does not correspond to the ones mentioned in Table 1. It was added at an intermediary stage, based on our knowledge of designations of account elements like "equity and investment fund shares" and "currency and deposits", and on frequency analysis of collocates in the corpus.

"_NN n.i.e.", [12] "_NN except * _NN" and "other _NN" may signal aggregates grouping heterogeneous things. For example, the patterns "_NN not included elsewhere" and "_NN n.i.e." indicate elements, which statisticians have not been able to record in any other category. Another example of a pattern signalling class names is "other _NN". Most term candidates matching that pattern were rejected. However, we kept those, which do not designate classes but concepts denoting entities. These are names of objects, which play a central role in the structuring of the BOP, like "other changes in financial assets and liabilities account" (designation of an account, in which all changes occurring during a period and not pertaining to transactions are recorded) and "other investment" (term denoting a functional category, which gathers specific investment relationships between residents and non-residents).

After this first selection and validation process, we obtained a shortlist of 148 term candidates in English.

4.2.3. Ontoterminology Design

The selected term candidates in English were all entered in the ontoterminology editor Tedi 3.7[13] [22]. Tedi is a software environment that allows creating multilingual ontoterminologies and exporting them into different formats (RDF, HTML, TBX, and CSV).[14] Based both on our understanding of the domain and on the list of term candidates, we defined in Tedi seven upper categories:

1. <**Entity**>: this category allows defining entities, whose activities are observed and analysed by statisticians for establishing the BOP.
2. <**Event**>: this category is the genus of all concepts representing activities and processes that lead to entries in the BOP.
3. <**Instrument**>: this category groups concepts representing instruments used by statisticians to collect, record, and present data on the BOP.
4. <**Location**>: this category models the residence of entities involved in a transaction, as relevant transactions mostly concern a resident and a non-resident entity. Interactions between entities that are resident in the same economy or that are both non-residents are, with very few exceptions, outside the scope of the BOP.
5. <**Principle**>: this category gathers concepts pertaining to principles, which statisticians have to adhere to when establishing the BOP. These principles are for example accounting rules or rules for data classification.
6. <**Product**>: this category encompasses concepts representing outcomes of production activities that are supplied or received by entities observed in the BOP, namely goods and services that are exported or imported.
7. <**Resource**>: this category is the genus of concepts representing objects used by entities to perform an economic activity. These objects are so-called economic assets. Those that are based on a financial contract, namely financial assets, can also be used by entities to perform an economic activity.

Each concept in the ontology was created as a combination of essential characteristics after one of these categories, and associated with a term in English.

[12] Abbreviation of "not included elsewhere".
[13] https://ontoterminology.com/tedi.
[14] https://ontoterminology.com/export.

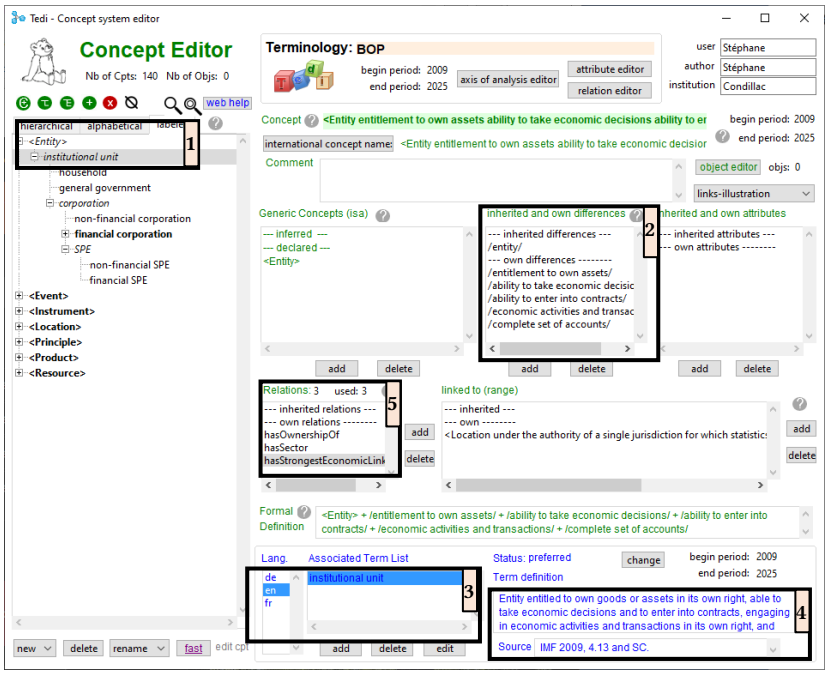

Figure 2: Definition of the concept <Institutional unit> in Tedi

Figure 2 shows the position of the concept <Institutional unit> as a specification of the category <Entity> (see rectangle [1] on the upper left of Figure 2), the combination of characteristics (specific differences) that defines it (2) and the term denoting it in English (3). A definition in English (4) was generated by concatenation of the term denoting the generic concept and of the labels in English of relevant specific differences. Nine domain-specific (associative)[6] relations were created, three of which can be seen in Figure 2 (rectangle 5).

Once each term in English was linked with a concept, we searched for terms that denote each concept in French and German. The respective corpus was used to validate known terms. As for unknown terms, we searched for knowledge-rich contexts (KRCs) [23] containing hypernyms or hyponyms of the terms to be found. Our assumption was that KRCs in French and in German containing hypernyms or hyponyms might supply hints for equivalents. This can be illustrated with a search for equivalents for the English terms “money market fund” (abridged “MMF”) and “non-MMF investment fund”. We knew that both terms are hyponyms of the term “investment fund”, and we were acquainted with equivalents for “investment fund” in French (“fonds de placement”) and in German (“Investmentfonds”). Queries designed to search for these equivalents respectively in the French and in the German subcorpus provided elements for equivalents of the hyponyms “MMF” and “non-MMF investment fund”. This was more straightforward for French than for German, though, due to limitations in the German subcorpus (see Section 5.4.3).

5. Discussion

5.1. Results

This research is still in progress. It has led so far to the creation of a trilingual ontology-based terminology resource on the BOP encompassing 140 concepts. Not all concepts of the domain have been analysed, nor all terms been extracted. The ontoterminology encompasses 150 terms in English, 151 in French, and 169 in German. In other words, the number of terms in each language is bigger than the number of concepts. Nevertheless, not all concepts were denoted by BOP terms: 10 concepts have no designation in any language. Among them are the seven upper categories (see Section 4.2.3): although they can be named in a natural language (e.g. “entity”, “event”...), they are not denoted by BOP terms. Still, they are necessary for the structuration of the concept system. Interestingly, the number of concepts without denotation is bigger in French (12) and in German

(20). We interpreted this as a confirmation that English is the language, in which the BOP has been conceptualised and standardised.

5.2. Content Validation by Experts

The ontoterminology was submitted to experts in order to assess whether the modelling work was reusable by others.[15] The validation of a terminology resource is a very important step because it confirms that the knowledge represented corresponds to a consensus among experts. Moreover, it should allow assessing the quality of definitions in natural language. To that end, we exported the ontoterminology in Tedi into two human readable formats: as an HTML electronic dictionary (see Figure 3) and as a concept map that can be edited in the software CmapTools[16] [24].

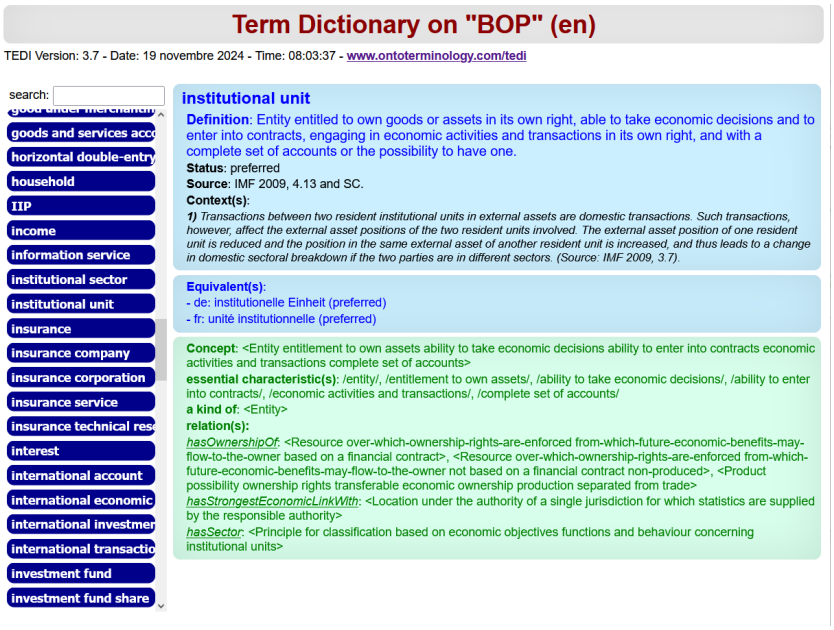

Figure 3: An entry in the HTML dictionary submitted to experts for validation

Experts gave a valuable feedback that allowed among others underlining modelling errors and discarding irrelevant concepts and terms. Modelling errors were the consequence of our limited knowledge of the BOP (as explained in Section 4.1) and of misinterpretation of corpus data. As for irrelevant concepts and terms, they resulted from the fact that the BOP is at the intersection of different disciplines (as mentioned in Section 3), and that we included in our corpus documents pertaining to neighbouring statistics, e.g. national accounts, because the BOP shares with these statistics some common concepts, and definitions are more precise in the reference manual for national accounts. We have not been able to reject straightforwardly those terms of national accounts that are not shared with international accounts and that are consequently irrelevant.

5.3. Ontology Validation by Competency Questions

After the validation by experts, a formal validation of the ontology with competency questions (CQs) [25] was performed. CQs are "natural language sentences that express patterns for types of question people want to be able to answer with the ontology. The ability to answer questions of the type indicated by a CQ meaningfully can be regarded as a functional requirement that must be satisfied by the ontology" [26]. In Tedi, the ontoterminology was converted into a knowledge graph in RDF format, which allows editing in Protégé.[17]

[15] The validation stage was limited due to time constraints.
[16] https://cmap.ihmc.us/.
[17] https://protege.stanford.edu/.

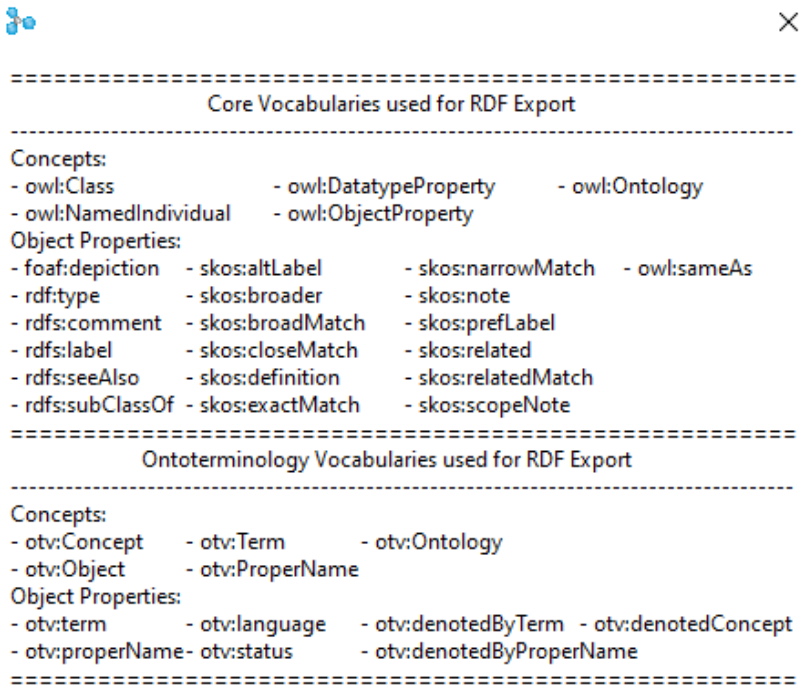

Figure 4: Vocabularies used in Tedi for the RDF export

Figure 4 is an excerpt of Tedi's help. It shows the different vocabularies used for the conversion into RDF. These include OWL, RDF, RDFS, SKOS, and OTV, a vocabulary conceived for the expression of essential characteristics as instances of classes.

The RDF knowledge graph was uploaded to the server http://www.ontologia.fr/OTB/BOP.rdf. The following CQs were defined:

1. **Designation of Concepts in Different Languages**: Which are the names of services relevant for the balance of payment in English, in French, and in German?
2. **Structuration of the Concept System by Generic Relations**: Which are, in English, the names of all economic assets that are based on a financial contract?
3. **Partitive Relations**: Which are the names of the different parts of the current account in English? Which are the names of the accounts making up the balance of payments in English?
4. **Associative Relations**: Can you designate in English all entities, which can own financial assets, non-produced non-financial assets or goods?
5. **Definitions**: What is the definition of a currency union? What is the difference between a customs union and a currency union?

All CQs were expressed with SPARQL syntax. BOP.rdf was queried through the SPARQL endpoint http://sparql.org/sparql.html. We present hereafter the expression of the second of the two CQs for partitive relations, namely "Which are the names of the accounts making up the balance of payments in English?" with SPARQL syntax.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX otv: <http://www.ontologia.fr/OTB/otv#>
PREFIX bop: <http://www.ontologia.fr/OTB/BOP#>

SELECT DISTINCT ?termEn

FROM <http://www.ontologia.fr/OTB/BOP.rdf>

WHERE {         ?cpt skos:prefLabel "balance of payments"@en.
                ?partCpt bop:partOf ?cpt.
          ?partCpt skos:prefLabel ?termEn.
  } ORDER BY ?termEn
```

The query searches for particular concepts in the BOP ontology (?partCpt bop), which are linked by the relation partOf with the concept denoted by “balance of payments” as its preferred term in English (skos:prefLabel). It returns the following answer, which is correct:

termEn
"capital account" @en
"current account" @en
"financial account" @en

5.4. Challenges

We mention in this section challenges that were faced in this research. Some are linked with the structuration of knowledge in the BOP (Section 5.4.1), others concern knowledge representation with the software environment chosen (Section 5.4.2), and a last category pertain to the size and composition of subcorpora (Section 5.4.3).

5.4.1. Structuration of Knowledge in the BOP

Not all knowledge units in the BOP can be defined straightforwardly by specific differentiation. This can be illustrated with the concept of residence. Entities (i.e. institutional units, see definition in Figure 3) that have their “centre of predominant economic interest” [4] in the same economic territory as the central bank or statistics agency establishing the BOP are regarded as resident. Those that have their strongest connection anywhere else in the world are considered non-resident. The concept of residence is fundamental because it determines whether an institutional unit will be considered for the establishment of the BOP of a given economy. In the modelling presented in this research, it has not been possible to represent individually the concepts of resident and non-resident because there was no relevant upper category, of which they could have been specifications in our ontology. Furthermore, being resident or non-resident cannot be considered as an essential characteristic, as an institutional unit can change its centre of predominant economic interest – and thus its residence – without becoming something different. But still statisticians have to determine the residence of the institutional units they observe. Finally, we decided to model the concept of residence as an instrument used by statisticians for the classification of institutional units. This decision was motivated by the fact that the residence of an

institutional unit cannot be defined independently from the central bank or statistics agency establishing the BOP.

Secondly, BOP compilers define broad categories that group different elements (entities, resources, products, etc.). These categories are relevant for the structuration of knowledge in the domain, because they reflect the way experts classify things. Nevertheless, they do not correspond to clear-cut concepts. They are easy to identify because they are clearly lexicalised, with linguistic means like "_NN n.i.e." or "other _NN" (see Section 4.2.2). We either split them into their component (e.g. "maintenance and repair services n.i.e." was split into "maintenance service" and "repair service"), rejected them and searched for the members of the class they designate (e.g. "other financial corporations") or kept designations that represent classes of objects that can be defined intensionally (e.g. "other investment").

Moreover, for certain concepts, the corpus supplied only extensional definitions, and it was not always unproblematic to determine the corresponding intension.

5.4.2. Knowledge Representation in the Software Environment Chosen

The ontoterminology created in Tedi can be exported into RDF, HTML, CSV, and TBX. RDF guarantees the interoperability of the resource on the Semantic Web [27] (see Section 5.3). Challenges were faced, among others because of missing data categories: at the time of writing, Tedi allows recording, for each term, a definition and one or more notes and contexts. However, it is not possible to record the source of notes and contexts in dedicated fields. We thus had to enter the information in the fields themselves.

As for the HTML dictionary, it provides very valuable conceptual information like the position of the concept in the concept system, its inherited and its specific differences, its genus, and the relations that may link it with other concepts (see lower part of Figure 3). That information is expressed in the formal language used by Tedi, with concept IDs that can be very long. The information is thus not immediately accessible for a human user. One possible improvement could be replacing each concept ID mentioned in that section with the term that denotes it – provided it is denoted in language, which is the case of most concepts in an ontoterminology.

To ensure interoperability with computer-assisted translation tools (CAT tools), the ontoterminology can be exported into TBX. However, the source of definitions is not included in the TBX export. Including a source for each piece of information in a terminology resource belongs to best practice in terminology management. This is why we regard this aspect in the TBX export as presenting room for improvement.

5.4.3. Size and Composition of Subcorpora

The small size of the subcorpus in German has limited the productivity of corpus analysis for that language. Whereas the size of the English subcorpus is 2 505 648 tokens (out of which 1 519 279 in reference documents and manuals [RMs]), and the subcorpus in French has 1 922 459 tokens (out of which 1 280 028 in RMs), the subcorpus in German encompasses only 633 046 tokens (out of which 217 591 in RMs). As a consequence, a certain number of terms in German could not be found in the corpus, and it was not possible to extract KRCs in German for most terms.

6. Conclusion

The BOP is a standardised domain. In that domain, concepts and terms are stable and based on a consensus among experts over a period of time (2009-2024 in the case of the research presented in this paper). We have acquired domain knowledge through the translation of texts, but no expertise. Moreover, we have direct access to experts (statisticians) establishing the BOP in our institution. Last, it was possible to compile a corpus of texts on the BOP in three languages, enabling us to complement our knowledge with authentic texts, i.e. "knowledge in action" [7]. All these elements justify the use of a methodology based on pre-existing knowledge, on corpus analysis, and on

interactions with experts. Our domain knowledge allowed searching the corpus more efficiently, as we had a broad idea of the content to be analysed. Furthermore, it made interactions with experts more fruitful. This makes this research original.

The RDF format chosen for knowledge graphs guarantees the interoperability of the resource created. The definition of a concept in a formal language is the basis for generating an expression in natural language of that definition. However, improvements can certainly be brought regarding interoperability with CAT tools and reusability by humans, be they translators or not.

This research is as mentioned still in progress. Challenges that were faced provide directions for future work. We believe that interviewing German-speaking experts will allow filling remaining gaps in the BOP terminology in that language. Additionally, the validation by experts shall be strengthened. Concerning knowledge modelling, while potentially challenging, modelling both concepts and classes of objects that structure knowledge in the BOP is necessary. The main reason for this is that experts do structure at least part of the domain with umbrella categories.

Additionally, it would be highly interesting to link the BOP ontoterminology with existing ontologies on neighbouring domains. Last but not least, a study of the diachronic dimension would be valuable, especially in view of the publication of the new BOP reference manual in March 2025.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] C. Laurén, J. Myking, and H. Picht, Terminologie Unter Der Lupe. Vom Grenzgebiet Zum Wissenschaftszweig. Vienna: TermNet, Internat. Network for Terminology, 1998.

[2] E. Wüster, Einführung in Die Allgemeine Terminologielehre Und Terminologische Lexikographie. 2nd ed. Fachsprachliches Zentrum, Handelshochschule Kopenhagen, 1985.

[3] R. Costa, 'Terminology and Specialised Lexicography: Two Complementary Domains.' Lexicographica 29(1):29–42. doi: 10.1515/lexi-2013-0004, 2013.

[4] International Monetary Fund, Balance of Payments and International Investment Position Manual, 6th Edition (BPM6). Washington, D.C., 2009.

[5] C. Roche, Le terme et le concept : fondements d'une ontoterminologie, in: Actes de la conférence TOTh 2007 : Terminologie & Ontologie : Théories et Applications, Annecy, France, 2007, pp. 1–22.

[6] International Organization for Standardization, ISO 1087-2019, Terminology Work and Terminology Science — Vocabulary. Geneva, Switzerland, 2019.

[7] C. Santos and R. Costa, Domain Specificity. Semasiological and Onomasiological Knowledge Representation, in: H. J. Kockaert and F. Steurs (Eds), Handbook of Terminology. Vol. 1, Amsterdam, 2015, Pp. 153–179, doi: 10.1075/hot.1.09dom1.

[8] International Organization for Standardization, ISO 704-2022, Terminology Work — Principles and Methods, Geneva, Switzerland, 2022.

[9] C. Roche, Should Terminology Principles Be Re-Examined? In: Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012). Madrid, Spain, 2012, Pp. 17–32.

[10] R. Studer, V. R. Benjamins, and D. Fensel, Knowledge Engineering: Principles and Methods, Data & Knowledge Engineering 25(1–2):161–98, 1998.

[11] International Monetary Fund. Balance of Payments and International Investment Position Compilation Guide. Companion Document to the Sixth Edition of the Balance of Payments and International Investment Position Manual. Washington, D.C., 2014.

[12] Bureau of Economic Analysis, U.S. International Economic Accounts: Concepts and Methods. Washington, D.C., 2024.

[13] European Commission. COMMISSION REGULATION (EU) No 555/2012 of 22 June 2012 Amending Regulation (EC) No 184/2005 of the European Parliament and of the Council on Community Statistics Concerning Balance of Payments, International Trade in Services and Foreign Direct Investment, as Regards the Update of Data Requirements and Definitions, 2012.
[14] European Communities, International Monetary Fund, Organisation Economic Co-operation, United Nations Development, and World Bank, System of National Accounts 2008 (SNA 2008). New York: United Nations, 2009.
[15] J. Pearson, Terms in Context. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.
[16] International Organization for Standardization, ISO 5078:2025(en) Management of Terminology Resources — Terminology Extraction, 2025.
[17] D. Maingueneau, Analyser les textes de communication. 3e ed. Paris, France: Armand Colin, 2016.
[18] P. Charaudeau, Dis-moi quel est ton corpus, je te dirai quelle est ta problématique, Corpus 8:37–66. doi: 10.4000/corpus.1674), 2009.
[19] A. Koester, Building Small Specialised Corpora, in: The Routledge Handbook of Corpus Linguistics. Abingdon: Routledge, 2010, Pp. 66–79.
[20] G. Doualan, De la représentativité à la spécialisation : exemple d'un petit corpus sur la synonymie, Corpus 18 | 2018. doi: 10.4000/corpus.3331.
[21] L. Anthony, AntConc [Computer Software], 2023.
[22] C. Roche, Tedi [Computer Software], 2024.
[23] I. Meyer, Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework, in: Recent Advances in Computational Terminology. Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2001, Pp. 279–302, doi: 10.1075/nlp.2.15mey.
[24] Institute for Human & Machine Cognition (IHMC), 'CmapTools [Computer Software]' 2019.
[25] M. Uschold, and M. Grüninger, Ontologies: Principles, Methods and Applications, The Knowledge Engineering Review 11(2):93–136, 1996.
[26] Y. Ren, A. Parvizi, C. Mellish, J. Z. Pan, K. van Deemter, and R. Stevens, Towards Competency Question-Driven Ontology Authoring, in V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai (Ed.) The Semantic Web: Trends and Challenges, Cham: Springer International Publishing, 2014, pp. 752–67. doi: 10.1007/978-3-319-07443-6_50.
[27] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web. A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities, Scientific American, May, 2001, 34–43.