Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities

Denis Maurel, Enza Morale, Nicolas Thouvenin, Patrice Ringot, Angel Turri

▶ **To cite this version:**

Denis Maurel, Enza Morale, Nicolas Thouvenin, Patrice Ringot, Angel Turri. Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities. Information, 2019, Natural Language Processing and Text Mining, 10 (5), pp.178. 10.3390/info10050178 . hal-02152978

HAL Id: hal-02152978
https://hal.science/hal-02152978v1

Submitted on 5 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

information

MDPI

Article

Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities

Denis Maurel [1,*], Enza Morale [2], Nicolas Thouvenin [2], Patrice Ringot [3] and Angel Turri [2]

1 Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT), Université de Tours, 37000 Tours, France
2 Institut de l'Information Scientifique et Technique, 54500 Nancy, France; enza.morale@inist.fr (E.M.); nicolas.thouvenin@inist.fr (N.T.); angel.turri@inist.fr (A.T.)
3 Lorraine Research Laboratory in Computer Science and Its Applications (Loria), Université de Lorraine, 54506 Nancy, France; patrice.ringot@loria.fr
* Correspondence: denis.maurel@univ-tours.fr

Received: 1 April 2019; Accepted: 18 May 2019; Published: 22 May 2019

Abstract: Istex is a database of twenty million full text scientific papers bought by the French Government for the use of academic libraries. Papers are usually searched for by the title, authors, keywords or possibly the abstract. To authorize new types of queries of Istex, we implemented a system of named entity recognition on all papers and we offer users the possibility to run searches on these entities. After the presentation of the French Istex project, we detail in this paper the named entity recognition with CasEN, a cascade of graphs, implemented on the Unitex Software. CasEN exists in French, but not in English. The first challenge was to build a new cascade in a short time. The results of its evaluation showed a good Precision measure, even if the Recall was not very good. The Precision was very important for this project to ensure it did not return unwanted papers by a query. The second challenge was the implementation of Unitex to parse around twenty millions of documents. We used a dockerized application. Finally, we explain also how to query the resulting Named entities in the Istex website.

Keywords: text mining; named entity recognition; data base of scientific papers; Istex; Unitex; CasEN; Docker

1. Introduction

1.1. Motivation

This paper describes the implementation of a mining tool which uses the named entities for the open access resource Istex, a database of twenty million scientific papers (https://www.istex.fr/). The main objective of the Istex Investment for the Future project (ANR-10-IDEX-0004-02) is to provide the whole of the French higher education and research community with online access to retrospective collections of scientific literature in all disciplines by driving a mass-scale national policy of document acquisition including journal archives, databases, text corpora and so on. The full text versions of these papers were bought by the French Government for the use of academic libraries. Papers are usually searched for using the title, authors, keywords or perhaps the abstract. To authorize new types of queries on Istex, we implemented a system of named entity recognition on all papers and we offer users the possibility to search using these entities.

We continue the Introduction section with a presentation of the French project Istex: First the beginning of the project (Section 1.2) and the current situation of the corpus (Section 1.3); Then, the expected service of Inist (Section 1.4) and the enrichment projects (Section 1.5). Section 2 speaks about the named entity recognition. After a presentation of the task (Section 2.1), we succinctly

describe the Unitex platform (Section 2.2). A first challenge was to build a new cascade in a short time (Section 2.3). The results of its evaluation showed a good precision, even if the recall was not very good (Section 2.4). The precision is very important for this project to prevent the return of unwanted papers by a query. The second challenge is the cascade implementation at Inist for the Istex project (Section 3). We wrote an annotation guide to precise what entities we wanted to annotate (Section 3.1), we organized the link between the document and the list of contained named entities with the use of a standoff file (Section 3.2). To run Unitex without fatal errors and without too much time consuming, we used a dockerized application (Section 3.3). Finally we describe the Istex system of queries to named entity enrichment (Section 3.4), the possibility to directly query the database (Section 3.4.1) or to build a complete TEI file (Section 3.4.2). We finally conclude in (Section 4) with some perspectives.

1.2. The Beginning of the Istex Project

The Istex Project was launched in 2012 and has documented resources acquired from around twenty publishers in the main scientific fields available to the French higher education and research community: Nearly 9000 scientific and technical journals representing a total of over 21 million articles, several thousand books, encyclopedias and dictionaries and some databases.

This acquisitions program ended on 1st January 2019 when the Ministry of Higher Education and Research's "Investments for the Future" program and its funding of around 60 million euros by the French National Research Agency both came to an end. In the future, other resources could be added to this corpus if a new acquisitions campaign is launched.

1.3. The Current Situation of the Corpus of Acquired Document Resources

The Istex corpus is still being constructed and has three specific characteristics: Firstly, it is mainly retrospective, although certain resources are very close to current publication dates (Wiley's for example). Secondly, it is fragmented because the program's budget did not provide for the acquisition of the whole scientific production of the world or even a homogeneous multidisciplinary set of documents, and thirdly, it is mainly made up of resources with subscription-based access. Open access resources which are sometimes essential (for example PLoS journals) are not included. The overall corpus is therefore not coherent in disciplinary terms if this dimension is taken into account.

Such gaps are normal in a corpus of this type. They are known and will be gradually filled by future acquisitions, by making arrangements with ongoing Istex acquisitions when negotiations make this possible and by opening Istex to quality scientific publications available in open access.

All members of the French higher education and research community can benefit from Istex. The exact definition is set out in Article 2 of the Licensing Agreement signed with publishers from whom the resources are acquired: "Licensees" shall mean legal entities under French law and situated in France on behalf of whom this Licence is subscribed herein, namely all government or private entities or organisations operating in higher education and research. All the contracts signed up until the present are based on this scope. It is however possible that future contracts do not include access for certain types of establishments, particularly public reading libraries. Currently, over 350 establishments have declared their IP addresses to access Istex resources.

1.4. The Services Provided by the Istex Platform

To respond positively to the project's strong ambitions, different technical challenges had to be solved. Firstly, the management of multiple flows of digital document acquisitions which are a consequence of the diversity of the publishers concerned, and the integration of a vast set of digital resources in a sole format which requires particularly rich and flexible analysis and standardization capacities. Secondly, the implementation of a very high performance document search system and the management and use of a large volume of textual annotations with a documentary architecture and the implementation of standardized formats, tools and protocols.

The Istex platform's technical development and hosting is carried out by the Institute for Scientific and Technical Information of the CNRS. The Istex platform enables users to access all the resources acquired from the different publishers via an API (https://api.istex.fr/documentation). From this point of view, it offers a greater level of user comfort than the various publishers' platforms. It can be integrated into and used on different existing sites, portals, tools and services (digital workspace, Google Scholar, document portals, etc.).

In the classic manner, the Istex API can be used to search for documents using a "Full-text" request type associated with different criteria. It thus provides access to document objects which represent a scientific article and are made up of several elements:

- The PDF file;
- The publisher's metadata;
- The standardized metadata in the MODS format;
- The XML display of the PDF in the TEI format;
- The document attachments (photos, graphs, etc.);
- The forms of enrichments.

The main use of Istex is a documentary one. All access to full text are recorded in the logs of the API. The consultation events are entered in the reporting dashboard manages by ezMesure (https://ezmesure.couperin.org/). Usage statistics concerning the access to full text are available on request from the French higher education and research laboratories. Usage statistics concerning the enrichment facets of Istex API are not currently available.

The notion of enrichment is a singular one in the Istex project. The acquisition, curation and standardization of publishers' documents can be used to carry out "texts and data mining" processing. These kinds of processing enrich the initial documents with complementary information.

Several explorations for the Text and Data Mining use of Istex are implemented. An example is the current work on the alignment of Unitex named entities <*placeName*> and <*geogName*> with the GeoNames database which allows you to locate the place recognized as a named entity. By using the identifier retrieved via GeoNames, you can access to additional information from Wikidata and DataBNF repositories (This alignment is available on https://placename-entity.data.istex.fr, https://geogname-entity.data.istex.fr and also on the SPARQL part of the website).

1.5. Enrichment

The need to enrich the documents acquired became rapidly apparent after the first documents were made available. The heterogeneity of the archive being built did not enable the creation of specialized corpora of documents based on specific selection criteria (a classification plan or transversal indexing). Several parallel work plans have been launched aimed at automatically classifying documents according to different classification plans, detecting and structuring the references cited and detecting several types of named entities.

The objective was to combine the use of state of the art tools and experimental tools derived from research to process a maximum amount of documents (several million) in the minimum time to eventually add new search criteria in the Istex API while providing access to the data produced. The volume of the Istex collection (over 21 million documents) led us to only use programs or algorithms capable of analyzing a document in less than a second to being able to process the maximum amount of documents in a reasonable duration.

Several partnerships have been set up to adapt programs or algorithms required for use on the Istex collections. The enrichment obtained with Unitex are the result of a close collaboration between the Inist-CNRS and the Lifat at the University of Tours.

2. Named Entity Recognition

In this section, we speak about the named entity recognition. After a presentation of the task (Section 2.1), we succinctly describe the Unitex platform (Section 2.2) and present our first challenge: To build a new cascade in a short time (Section 2.3). We end this section with an evaluation (Section 2.4).

2.1. Presentation

Named Entity Recognition (NER) appeared with the MUC Conferences (Message Understanding Conferences). The MUC challenge was information retrieval, with some factual questions about a corpus news such as "Which terrorist group planted a bomb?", "Where?", "How many deaths?" or "Which factory bought another one?" and so on. So the organizers defined Named Entity as the persons' names, location names, organization names, dates, percentages, currency [1]. Other evaluation campaigns sometimes added titles, hours, roles and so on. A state of the art of NER can be found in [2,3].

Most of the NER systems use machine learning techniques. But machine learning is not really an automatic process, because of the preliminary work of the annotated corpus with the inherent difficulty of annotation and the inter-annotator agreement [4]. For instance, ref. [5] recognized Named Entities in English, Spanish, Dutch and Czech with CRFs; ref. [6] used artificial neural networks; and the multilingual system of [7] treated the text as bytes flow. The task is now centered on other texts as social networks, for instance [8] or entity linking from text to Semantic Web [9–11] or geographical entities linked to map [12].

The main idea is to use internal and external evidence [13], i.e., the local context. For instance, the sequence *Charles de Gaulle* is recognized as a person's name, because of the first name *Charles* - internal information - and the sequence *General de Gaulle* is also recognized as a person's name, because of the title *General* - external information -. The NER is possible with three approaches, machine learning, symbolic rules or hybrid approaches. Machine learning techniques need a training corpus which was not available for this task and too time-consuming because of the great heterogeneity of the corpus (physical or biological sciences, mathematics, human sciences, history, geography, education and so on. Therefore we used an approach based on symbolic rules [14,15], or more precisely on a cascade of rules processed by Unitex Software (see Section 2.2). The use of local context is in adequation with Unitex graph descriptions.

2.2. Unitex Software

Unitex software is a free open source software which analyzes textual data (https://unitexgramlab.org/). It is high-performance in computing terms and has a user-friendly interface with language resources distributed out-of-the-box.

The Unitex Language Processing engine is based on automata-oriented technology similar to the augmented transition network (ATN) model. For instance, it particularly allows users to:

- Insert, move or replace characters on the text processed;
- Compile rules and dictionaries as Finite-State Machines;
- Use variables instanced with a part of the text or with any characters;
- Work into a sequence of letters;
- Use regular expression;
- Build cascades of rules.

The Visual Integrated Development Environment of Unitex enables users to easily build a project even without computing experience, create specific dictionaries, test graph-based rules and, of course, to apply all of these resources to text files. The project also can integrate multilingual resources shared on the website by a large community of users with 23 languages. When the rules are tested, we can begin faster production with specific scripts.

Figure 1 presents the graph to encapsulate XML tags. Unitex parses the text from the first box on the left to the last box on the right. Gray boxes contain a call to a sub graph. The output is in bold characters under boxes. If we consider as an example the text *<publisher>Springer-Verlag</publisher>* extracted from the header of a paper, Unitex follows the path:

```
(1): The first box recognizes "<" and merges "{" before "<publisher>";
(2): The second box initialize the variable name with "publisher";
(3): The third box recognizes ">";
(4): The fourth box merges ",.XmlTag+name" after "<publisher>";
(5): The fifth box tests if the variable nameSpace is instancied;
(6): The sixth box concatenates "Bpublisher+BeginTag+grftoolXml}"
to the precedent merging.
```

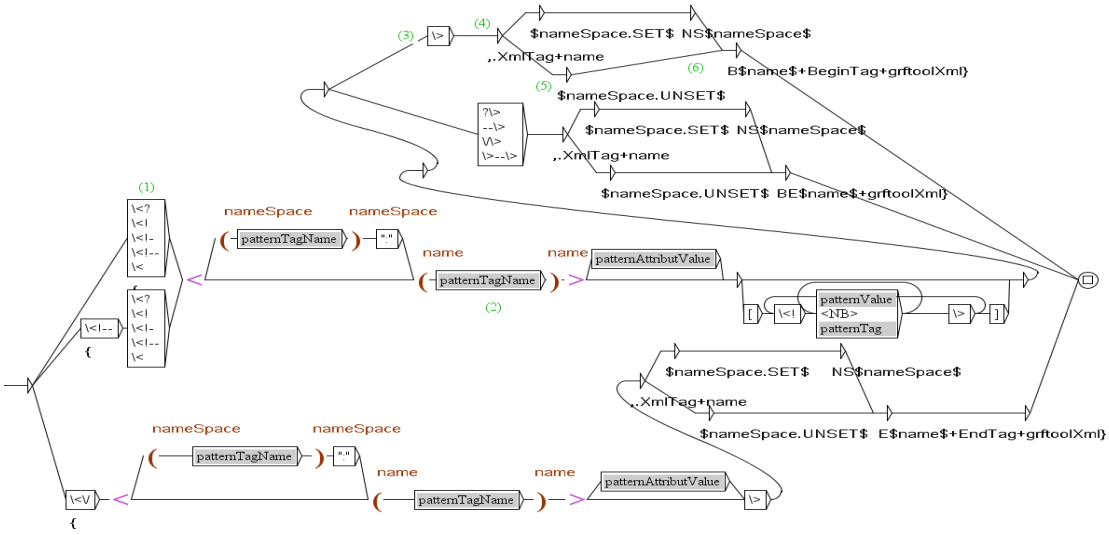

Figure 1. The graph to encapsulate XML tags.

So the output is merged into the text to give:

```
{<publisher>,.XmlTag+namePublisher+BeginTag+grftoolXml}
Springer-Verlag
{</publisher>,.XmlTag+nameEpublisher+EndTag+grftoolXml}
```

Unitex allows the use of graph cascade with the CasSys program inspired from [16]. Cascades [17] are used in many NLP applications, as chunking [18], syntactic analysis [19], morphological analysis [20] and so on.

The graph shown in Figure 1 is the first in our cascade. The principle of a cascade is simple: Each graph of the cascade parses the output text of the preceding graph. The graph order is difficult to decide but it is very important for a good parsing. For instance, you have to recognize a building name like *Charles de Gaulle airport* after you recognized *Charles de Gaulle* as a person's name. Another example is shown in Figure 2: to encapsulate URL, you have to use the XML tags *ext-link*, if these exist in the paper (1) or you have to describe the URL (2).

The distribution of Unitex - and its resources - is free under the terms of the Lesser General Public License (LGPL).

Before the use of Unitex in the Istex project, the software was only used to process some relatively homogeneous collections (often newspapers) of a middling quantity of texts. The new challenge of Istex was to make Unitex robust even when processing different types of academic texts (biology, mathematics, history, geography, human sciences and so on) in a very large collection of almost twenty million documents. This point is presented in Section 3.3.

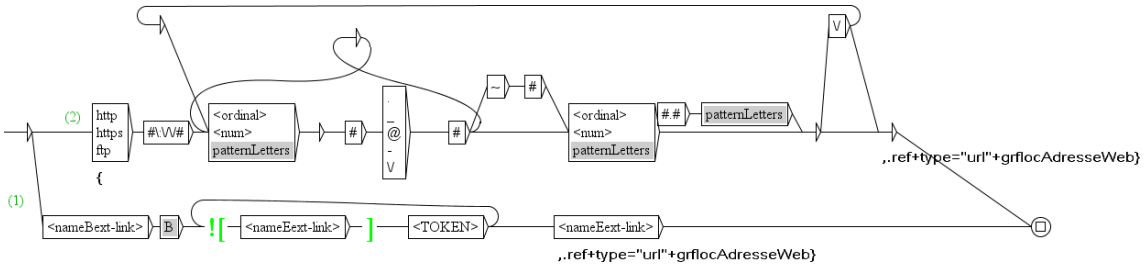

Figure 2. The graph to encapsulate URL.

2.3. NER with Unitex

For more than a decade, the Lifat has been developing cascade NER rules for French, CasEN, which is a free and open source, under a LGPL licence (http://tln.lifat.univ-tours.fr/Tln_CasEN.html). After adjudication, CasEN probably was the first software in the *Etape* campaign (http://www.afcp-parole.org/spip.php?rubrique132). So the first point in the Istex project was to try CasEN on the section of the Istex collection written in French. We worked on the improvement of CasEN for scientific texts in French, and in parallel built a new cascade for scientific texts in English. The French version of CasEN is based on very different corpora (newspapers, querying, books, scientific papers...) but the English version is only based on the Istex collection, i.e., scientific papers.

In this section, we describe the use of the French cascade CasEN. The use of the English one is similar but less complete. CasEN analyzes text in four steps:

1. Preprocessing
   - We research the beginning of the bibliography (i.e., the end of the text to analyze);
   - We normalize the text (spaces, tabulations and line feeds; apostrophes, quotation marks, hyphens and ellipses);
   - We tokenize the text (sequences of letters or single other characters);
   - We apply dictionaries (common words, proper names and specific CasEN dictionaries).
2. Analysis: We apply the first subcascade in the defined order
   - 10 tool graphs (XML tags, numbers, specific multiword units);
   - Four amount graphs, 12 time graphs;
   - Six person graphs, two product graphs, five organization graphs;
   - Four location graphs, two event graphs, role graph, address graph;
   - A reference graph, funder and provider graph (organizations with a specific role in the paper).
3. Synthesis (the second subcascade)
   - We transform the XML-CasSys file into a TEI file; (https://tei-c.org/)
   - We customize it according to the target guide (Section 3.1).
4. Counter (the third subcascade)
   - We define the tags that we want to list and count;
   - We build a standoff file (Section 3.2).

Figure 3 presents the skeleton of a cascade of Unitex graphs. We start with the preprocessed text and the first graph (the graph of Figure 1) modifies it. Now the XML tags are considered as multiword units with the category *XmlTag*. We parse this modified text with the second graph (the graph of Figure 2) that replaces the URLs by multiword units with the category *ref*. This continues until the last graph of the cascade, which produces the final text.

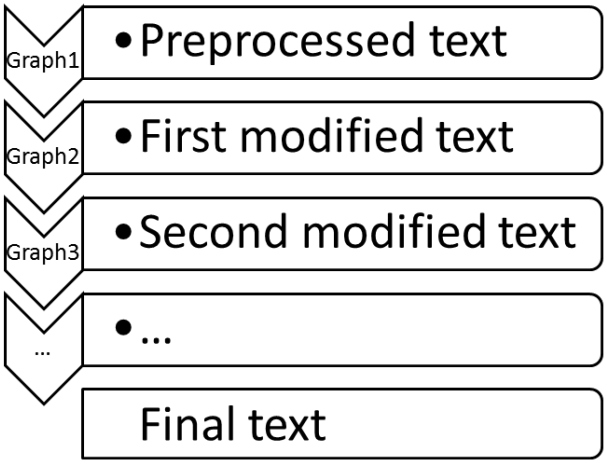

Figure 3. Cascade of graphs.

2.4. Evaluation

2.4.1. Evaluation Procedure

The purpose is to evaluate the English cascade build for the Istex project by the Lifat at the University of Tours (France). An English corpus of 49 Istex documents was manually annotated with the ten named entities used in the Istex project (This English corpus is available on https://unitex-anglais.corpus.istex.fr/). This annotated corpus is then compared with the same corpus tagged with Unitex (Istex linguistic package).

The first part of the evaluation consists of finding and counting properly named entities correctly detected and different types of errors in the tagging of the documents. As example, we consider the following sentence: *Adult sheep tissues were obtained from non-pregnant ewes from the University of Tasmania Animal Farm*, in which there is only one named entity: *<orgName>University of Tasmania Animal Farm</orgName>*.

Three different types of tagging errors may be encountered. The first type is that which concerns the location of the tags. If the tags are misplaced, it is a boundary error: *the <orgName>University of Tasmania</orgName> Animal Farm*. The second type is that which concerns the labelling of the tags. If the label of the tag is not the expected one, it is a typing error: *the <placeName>University of Tasmania Animal Farm</placeName>*. The third type is a combination of the two preceding errors, a boundary and typing error. Two other cases must be recorded in the calculated data for the evaluation. The first case is when a neutral term is recognized as a named entity: *<orgName>Adult sheep tissues</orgName>*. It is an insertion. The second case is when a named entity is not recognized. It is a suppression.

2.4.2. Results

The second part of the evaluation consists of calculating the classical measures of Recall and Precision, and also the Slot Error Rate (SER) [21]. Table 1 explains the calculated data.

Table 1. The calculated data.

I = entities detected by mistake (insertion)	Slot Error Rate (SER): D + I + TE + 0.5 (T+E) / R
D = entities totally missed (suppression)	Recall: (S-I)/R
T = incorrect typing	Precision: (S-I)/S
E = incorrect boundary	Typing accuracy: (S-I-T-TE)/S
S = detected entities	Tagging accuracy: (S-I-E-TE)/S
R = real entities	

The founded results on the English corpus of 49 Istex documents manually annotated for the test (see Section 2.4.1) are detailed in Table 2.

Table 2. The results in the 49 annotated Istex documents.

D	I	T	E	TE	S	R
1516	281	64	265	131	3296	5414

Finally, Table 3 presents the result measures:

Table 3. The result measures.

SER	Recall	Precision	Typing Accuracy	Tagging Accuracy
38.6%	55.7%	91.5%	85.6%	79.5

The Recall is weak, because the English cascade was not finished at the end of the project (for the evaluation time). As we said in Section 2.3, we developed CasEN before the Istex project, but only for French. The evaluation described here concerns the English version of CasEN. We are always working this cascade. In this project, the most important point is the precision, because one doesn't want to find a text with an error of tags when one queries the database (Section 3.4). The difficulty is the extreme diversity of the Istex documents: Mathematics, physics, astronomy, biology, psychology, history, geography, and so on.

3. NER Implementation at Inist

This section speak about our second challenge: The cascade implementation at Inist for the Istex project. We wrote an annotation guide to precise what entities we wanted to annotate (Section 3.1), we organized the link between the document and the list of contained named entities with the use of a standoff file (Section 3.2). To run Unitex without fatal errors and without too much time consuming, we used a dockerized application (Section 3.3). Finally we describe the Istex system of queries to the named entity enrichment (Section 3.4), the possibility to directly query the database (Section 3.4.1) or to build a complete TEI file (Section 3.4.2).

3.1. NER but for Which Entities?

At the beginning of the Istex Project, 10 entities to be detected were chosen by the initial workshop team. They are basic entities for some of them and more specific for others:

- Names of persons: <persName>

```
We thank Prof. <persName>Harry Green</persName> and
Dr <persName>Larissa Dobrzhinetskaya</persName> for assistance
```

- Administrative place names: <placeName>

```
in northern <placeName>Sweden</placeName>
```

- Geographical place names: <geogName>

```
located on the coast of the <geogName>Baltic Sea</geogName>,
on the inlet to <geogName>Lake Malaren</geogName>
```

- Dates (year): <date>

```
Friday, 14 May <date>1993</date>
```

- Names of organizations: <orgName>

```
published by the <orgName>National Bank of Belgium</orgName>
```

- Funding organizations and funded projects: <orgName type= "funder">

```
This work was supported by
<orgName type="funder">INCO-DC grant IC18CT96-0116</orgName>
from the <orgName>European Commission</orgName>
```

- Provider organizations of resources: <orgName type="provider">

```
numerical simulations were run on
the <orgName type="provider">PARADOX-III cluster</orgName> hosted by [...]
```

- URL: <ref type="url">

```
a community web site <ref type="url">www.scruminresearch.org</ref>
is being built.
```

- Pointers to bibliographic references: <ref type="bibl">

```
proposed by <ref type="bibl">[Broyden12]</ref>
```

- Bibliographic references: <bibl>

```
«the scene of our finitude, the place where we encounter
the limits of our subjectivity».
<bibl>Diane Michelfelder et Richard Palmer, Dialogue and Déconstruction.
The Gadamer-Derrida Encounter, Albany, Suny Press, 1989, p. XI</bibl>
```

The annotation guide of Named Entities Istex Project is available at http://tln.lifat.univ-tours.fr/Tln_CasEN_eng.html. The ISTEX Project uses a light TEI: The labelling does not include detail tags. For instance, *Franklin Delano Roosevelt* is tagged in TEI: *<persName><forename>Franklin</forename> <forename>Delano</forename> <surname>Roosevelt</surname></persName>*; and in the Istex project: *<persName>Franklin Delano Roosevelt</persName>*.

3.2. The Use of a Standoff File

The enrichment available on the Istex API are presented by means of an external XML standoff. This choice for Istex enrichments was based on the opportunity to use several tools for tagging or categorizing the full-text in Istex documents. Then in order to separate the enrichments, the Istex API offers a standoff for each tool used to obtain enrichment.

In the Unitex standoff, the named entities are tagged according to the standard TEI XML (Text Encoding Initiative) which allows the exchange of textual data stored digitally by users of various computer systems.

The standoff is composed of:

- A <teiHeader> containing a <fileDesc>, a <encodingDesc> and a <revisionDesc>;
  - The <fileDesc> (1) contains information about the type of enrichment achieved and the conditions for using the standoff data;
  - The <encodingDesc> (2) indicates the name of the application used to obtain the enrichment;
  - The <revisionDesc> (3) indicates the date of Unitex version change and the name of the new version;
- A series of <listAnnotation> (4), one for each type of named entity recognized; the <listAnnotation> contains an <annotationBlock> for each named entity belonging to this type;
  - An <annotationBlock> contains one named entity and information about this term; the <numeric value> indicates how often this named entity occurs in the document.

```
(1)
<fileDesc>
    <titleStmt>
        <title>Reconnaissance d'entités nommées</title>
            <respStmt>
                <resp>enrichissement entités nommées Istex</resp>
                <name resp="Istex-rd">Istex</name>
            </respStmt>
    </titleStmt>
    <publicationStmt>
        <authority>Inist-CNRS</authority>
        <availability status="restricted">
            <licence target= "http://creativecommons.org/licenses/by/4.0/">
                <p>L'élément standoff de ce document est distribué sous
                    licence Creative Commons 4.0 non transposée (CC BY 4.0)</p>
                <p>Ce standoff a été créé dans le cadre du projet Istex -
                    Initiative d'Excellence en Information Scientifique
                    et Technique</p>
            </licence>
        </availability>
    </publicationStmt>
    <sourceDesc>
        <biblStruct>
            <idno type="Istex">33B1ACBC65792C01FDFEA7F39C33CE61D57D2FCB</idno>
        </biblStruct>
    </sourceDesc>
</fileDesc>

(2)
<encodingDesc>
    <appInfo>
        <application ident="UnitexCasSys" version="{VERSION}">
            <label>Unitex CasSys</label>
        </application>
    </appInfo>
</encodingDesc>

(3)
<revisionDesc>
    <change who="#Istex-rd" when="2017-03-16"
        xml:id="unitex-3.2.0-alpha">version 2830</change>
</revisionDesc>

(4)
<listAnnotation type="orgName" xml:lang="en">
    <annotationBlock corresp="text">
        <orgName change="#Unitex-3.2.0-alpha" resp="Istex-rd"
            scheme="https://orgname-entity.data.Istex.fr"/>
        <term>NSW Department of Environment and Conservation</term>
        <fs type="statistics">
            <f name="frequency"><numeric value="1"/></f>
        </fs>
```

```
    </annotationBlock>
</listAnnotation>
<listAnnotation type="persName" xml:lang="en">
    <annotationBlock corresp="text">
        <persName change="#Unitex-3.2.0-alpha" resp="Istex-rd"
            scheme="https://persname-entity.data.Istex.fr"/>
        <term>D. W. Larson</term>
            <fs type="statistics">
                <f name="frequency"><numeric value="2"/></f>
            </fs>
    </annotationBlock>
    <annotationBlock corresp="text">
        <persName change="#Unitex-3.2.0-alpha" resp="Istex-rd"
        scheme="https://persname-entity.data.Istex.fr"/>
        <term>R. Shine</term>
        <fs type="statistics">
            <f name="frequency"><numeric value="15"/></f>
        </fs>
    </annotationBlock>
</listAnnotation>
```

3.3. Unitex Implementation at Inist

The Istex-RD team decided to delegate the processing of Unitex/CasEN on the Istex corpus to the IT Operation team of Inist (Iprod). The first goal was to evaluate the CasEN accuracy on a small bilingual (FR, EN) corpus, then to strengthen the process by scaling up to more than 2 millions of documents, and finally to apply it to the Istex corpus.

Iprod was associated early in this project, whereas the software and linguistic components involved were still in development mode because the Lifat was working on the CasEN cascades for French and English. In parallel, in order to process a 20 million document corpus, the Ergonotics start-up developed an add-on for Unitex to enhance Unitex overall performance, as well as a custom Unitex build script taking this add-on into account. The Unitex software itself was regularly improved using the feedback resulting from the first results of processing Unitex/CasEN on Istex documents (https://github.com/UnitexGramLab/ - the fact that one of the Ergonotics founders is an active committer of Unitex was a major point in this continuous improvement process).

To implement this Unitex development workflow at Inist, Iprod had to build and test frequently the processing chain, integrating the above-mentioned artefacts which were delivered at a different pace by each partner. It was also a requirement for Iprod to be able to run the processing chain using different versions of the tool chain (gcc/g++) used to build the Unitex software. After some iterations on increasing workloads, we concluded that the optimizations resulting from using a tool chain instead of another where not significant enough compared to the time required by the linguistic part of the process.

For the sake of operational stability, Iprod works on a limited number of operating systems (OS) as the basis for its production environment (SLES 11 and Ubuntu 12, 14 LTS at that time). Using different tool chains not available on the shelf in the different supported Linux distributions was a problem for an operation team which needed to constantly limit its technical landscape in order to maintain the same level of service quality.

Docker (https:/www.docker.com/) was still in its early phase in 2015, but was the technical solution Iprod used to stick to its principles and to experiment in an agile way. Docker (which is a Linux technology) is a containerization solution. It allows the user to embed its application (e.g., Unitex, CasEN) and its running dependencies (e.g., the required part of an OS in a precise version) into

an image, which can be used everywhere Docker is installed. Most of the time, a docker image is created using a specialized language (Dockerfile). This automated way of building operational environments for applications greatly contributes to increase their operational lifetime, their reusability and their maintainability.

To run a dockerized application, it suffices to execute the corresponding image using a docker run command which creates a Linux container to run the application. The container is no more than an ordinary process but more isolated than usual from the rest of the OS. It is up to the user to define which parts (network, filesystem, process, etc.) of the host OS (called the Docker host) are shared with the running container.

One of the great advantages of using containerization solutions like Docker is that they enable users to uncouple what has to be executed (e.g., software running only on Cent OS 7) from the location they have to run (e.g., Debian 9), provided that the location is running under a Docker compatible OS (Linux but also Mac OS and Windows using user transparent Linux Virtual Machines), and that the embedded software is not compiled for a specific CPU instruction set which is not universally available.

Cloud and Docker are often associated but some scientific communities quickly realized that Docker could be used as a convenient way to package and deploy "not so easy to build" software on different locations ranging from another user workstation up to grid environments (https:/biocontainers.pro/). The fact that other containerization solutions are available apart from Docker and Singularity is worth mentioning. Singularity is more recent than Docker (2015 vs. 2013), fits better in the High-Performance Computing (HPC) world as it does not require special privileges for the user needing to run a container (this is a no-go in highly shared environments), and integrates better in the HPC ecosystem [22].

The main advantages of using Docker in the Unitex/CasEN project were the following: Iprod was able to use technical environments normally unsupported (ahead of time) in its production environment and Docker offered a convenient way to automate the build of the software stack made up of the assembly of Unitex, Ergonotics add-ons and CasEN. We can add the possibility of building easily different assembly combinations (e.g., different versions of Unitex associated with different versions of CasEN, built using different toolchains), the opportunity of using these different (hard to build) combinations in different locations (development workstation, server) and the ability to maintain the Unitex/CasEN workflow over the years (as Docker helps the user to describe fully its building process from scratch).

The way the resulting Docker images were built was influenced by the different artefacts to assemble, their origins and their different delivery rates.

- Step 1: Unitex compilation
  - Parameters
    - The Ubuntu version number to use (implicitly the gcc tool chain to use for the build process);
    - The Unitex version number to use (it should be noted here that the Unitex software itself was updated during the process to fix bugs or to improve things related to this experiment);
    - A zip files from Ergonotics containing an optimization add-on and a custom Unitex build script.
  - Output
    - A Docker image named unitex/compiled whose version number (tag in Docker lingo) is a number made up of the version number of Ubuntu and the version number of Unitex (e.g., unitex/compiled:14.04_2903).
- Step 2: Assembly of Unitex and CasEN
  - Parameters

    - The previously built unitex/compiled Docker Image to start from;
    - The CasEN artefact to use (the version number in this case is a date).
- Output
    - A docker image named unitex/runable whose tag is a number made up of the unitex/compiled tag and the version number of the CasEN artefact (e.g., unitex/runable: 14_04_2903_20151201).

This two-step construction process had the advantage of saving the build time when assembling a different version of CasEN with the same Unitex base image, allowing us to effectively compare their outcomes or to highlight regressions cases.

Without going into details, this build process was designed at the beginning of 2015 and would benefit from being improved. The reason is that since then, Docker has added a lot of new functionalities (https://docs.docker.com/develop/develop-images/multistage-build/), particularly the multistage build, as well as a smart solution to take corporate proxies into account without giving the proxy knowledge to the docker image (https://docs.docker.com/network/proxy/). Also it would be possible to use Singularity instead of Docker (The interested reader can refer to https://github.com/ecirtap/unibat/ to have a look at the software stack that was used to setup this build process).

The unitex/runable image uses four parameters: The number of threads to use to apply CasEN to different documents in parallel, the directory containing the documents to process, the directory which will contain the results (standoffs) and the name of the CasEN script to use (there is a script for the English and French cascades in each of the CasEN artefacts).

The only shared resources between the running container and the host are the two directories mentioned above. A report is produced in real time during computation, including the version of the different components used, the time required to compute the standoff for each document and the output of Unitex cascades. As the computation is multithreaded, if one of the processed documents generates a segmentation fault, all the documents processed at the same time are aborted.

We applied this workflow to a corpus of documents of increased sizes (1 K, 10 K, 1–2 M, etc.) to detect blocking bugs as soon as possible, and thus strengthen the Unitex/CasEN integration for Istex. To this end Iprod used a Virtual Machine with 32 vCPU and 12GB of memory (the underlying hardware was, depending on the time of the experiment, two-socket Xeon E5-2660 Dell R620/R630 servers ranging from eight to 14 cores per socket running under ESXi v5.5).

At the end of the may 2016 campaign, 2,385,436 documents was submitted using a batch of approximately 9320 documents to a 20 thread Unitex/CasEN container and 2,349,421 standoffs was generated. With an average of 4.21 documents processed per second for a total of elapsed time 553,057 s (6.5 days).

2% of the remaining documents were left unprocessed in their vast majority because they were part of a batch containing a document producing a segmentation fault. For each faulty batch, a maximum of 20 documents were transmitted for further examination to Ergonotics and LIFAT, which in return produced fixes which were used to produce a new Unitex/runable docker image which was applied to the unprocessed part of the corpus.

Following this strengthening campaign, the Unitex/CasEN workflow was applied to 15,847,345 documents of the Istex corpus, all of which were accessible to readers using the Istex demonstrator website with the faceting functionality (enrichment type) (http://demo.istex.fr/).

3.4. Named Entities Queries in the Istex Website

A query on the ISTEX API consists of:

- The basic URL: *https://api.istex.fr/document/?*
- A mandatory parameter: *q={query}*
- Optional parameters:

  - *output={list of fields to display}*
  - *size={maximum number of documents displayed}*
  - *from={number of the first document}*
- A parameter separator: &

The query documentation is available at https://doc.istex.fr/tdm/requetage/ for more information about the syntax of queries (Lucene query language). We will give in this section three examples of a query.

The first one is a very simple query: search for the first document containing the term "forestry" in the Istex database:

```
https://api.istex.fr/document/?q=forestry&size=1
```

The system finds 707,274 documents in the ISTEX database contain this term. The results show the title and identifier of the first one.


```
{
  "total": 707274,
  "nextPageURI":
    "https://api.istex.fr/document/?q=forestry&size=1&defaultOperator=OR&from=1",
  "firstPageURI":
    "https://api.istex.fr/document/?q=forestry&size=1&defaultOperator=OR&from=0",
  "lastPageURI":
    "https://api.istex.fr/document/?q=forestry&size=1&defaultOperator=OR&from=9999",
  "hits": [
    {
      "arkIstex": "ark:/67375/56L-3RD0LK4K-P",
      "title": "Windsor-Forest. To the Right Honourable.
        George Lord Lansdown. By Mr. Pope.",
      "id": "10A93EC345EC7D19CBAE42E44ABD13CED8DACD0E",
      "score": 14.641167
    }
  ]
}
```


The second one is the search for documents from the journal with the title "Biofutur", whose ISSN is "0294-3506", which were published in 1955 and whose author is "Dodet". We wish to visualize the title, the author and journal information for the first 100 documents:

```
https://api.istex.fr/document/?q=(host.title:"Biofutur"+OR+host.issn:"0294-3506")
    +AND+host.publicationDate:1995+AND+author.name:"DODET"
    &output=title,author,host&size=100
```

The result is just one document.

The third one allows to obtain the documents containing a Unitex enrichment with the *persName Chomsky*:

```
https://api.istex.fr/document/?q=namedEntities.unitex.persName:"chomsky"
```

The result is 588 documents, but the default result of this query is the title and identifier of the first 10 documents containing the <*persName*> *Chomsky* in a Unitex enrichment. It is possible to obtain titles and identifiers for the 588 documents with the modified query:

```
https://api.istex.fr/document/?q=namedEntities.unitex.persName:"chomsky"&size=588
```

3.4.1. A More Complex Example of Query Concerning Unitex Enrichment

The following query finds out the number of documents in Istex database containing Unitex enrichment:

```
https://api.istex.fr/document/?q=*%20AND%20enrichments.type.raw:(%22unitex%22)
    &facet=corpusName[*]&size=10&rankBy=qualityOverRelevance&output=*&stats
```

We can obtain this complex query using the facets available on the Istex demonstrator: Click on "Continuer", write a "*" as a keyword, choose the facet "Types d'enrichissement" and click on "unitex" (see Figure 4). Then copy the query, and paste it into your browser.

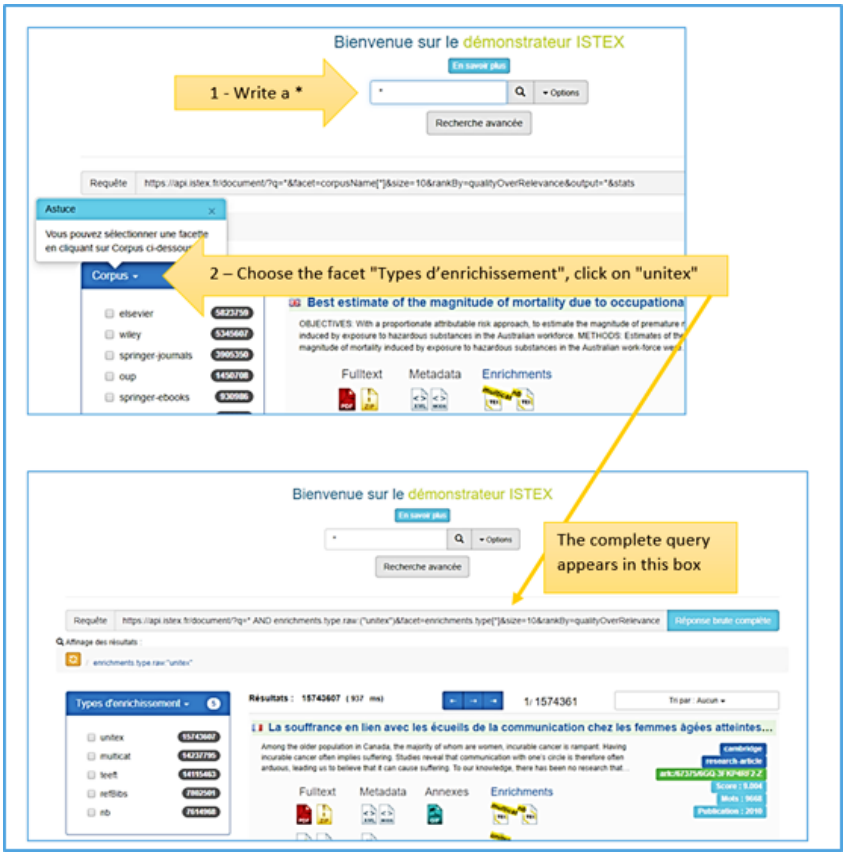

Figure 4. Procedure for obtaining the complex query to find out the number of documents containing Unitex enrichment.

The results of this query show that 15,743,607 documents on the Istex API contain a Unitex enrichment, on 30 April 2019.

3.4.2. Replace Unitex Enrichment in the Complete TEI File of a Document

As we have seen in Section 3.2, each enrichment for a document on Istex API is presented in an external TEI tag: A standoff. The aim of using a standoff is to allow the user to look at each enrichment separately. The query to find out the named entities recognized by Unitex for a document (in a standoff), using the Istex identifier of the document *697E812020AD421C96073D118759E7525A9E7DE2* is:

```
https://api.istex.fr/document/
697E812020AD421C96073D118759E7525A9E7DE2/enrichments/unitex
```

The complete TEI file of the document which contains metadata, full text and bibliographic references can be completed by this standoff. We have just to add *?consolidate* at the end of the query:

```
https://api.istex.fr/document/
697E812020AD421C96073D118759E7525A9E7DE2/enrichments/unitex?consolidate
```

4. Conclusions

We have presented the Istex project enrichment by queries on named entities contained in scientific papers. This service needs robust algorithms to recognize named entities in a very large and heterogeneous corpus of twenty million scientific papers. The main scientific contributions of this project were the improvement of the linguistic platform Unitex and the adaptation of the NER system CasEN to this particular corpus. We presented the evaluation of CasEN with 91.5% precision, the organization of a standoff file for results and the implementation with a dockerized chain of treatment.

We plan in the future to complete the cascades and increase the recall, to correct the graphs and increase the accuracy. We will compute the periods written as *1985–1992* or *the 90s* instead of only taking into account actual years. Technically, we project to update our processing chain (by using *Singularity* instead of *Docker*) and scale it up horizontally in order to speed up the whole process, distributing it on multiple nodes.

Author Contributions: Supervision, D.M.; Writing—original draft, D.M., E.M., N.T., P.R. and A.T.

Funding: This research was funded by the French government project ANR-10-IDEX-0004-02.

Acknowledgments: The authors thank Julien Franck, Anubhav Gupta and Sevil Zeynali for their contributions to the project.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Chinchor, N. Muc-7 Named Entity Task Definition. 1997. Available online: https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html (accessed on 21 May 2019).
2. Nadeau, N.; Sekine, S. A survey of named entity recognition and classification. In *Named Entities: Recognitionandclassification and Use*; Sekine, S., Ranchhod, E., Eds.; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2009; pp. 3–28.
3. Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2145–2158.
4. Fort, K. Les ressources annotées, un enjeu pour l'analyse de contenu: Vers une méthodologie de l'annotation manuelle de corpus. Ph.D. Thesis, Université Paris-Nord-Paris XIII, Paris, France, 2012.
5. Konkol, M.; Brychcín, T.; Konopík, M. Latent semantics in named entity recognition. *Expert Syst. Appl.* **2015**, *42*, 3470–3479. [CrossRef]
6. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
7. Gillick, D.; Brunk, C.; Vinyals, O.; Subramanya, A. Multilingual Language Processing from Bytes. *arXiv* **2015**, arXiv:1512.00103.
8. Ritter, A.; Clark, S.; Etzioni, M.; Etzioni, O. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Stroudsburg, PA, USA, 27–31 July 2011; pp. 1524–1534.
9. Cano, A.; Rizzo, G.; Varga, A.; Rowe, M.; Stankovic, M.; Dadzie, A.S. Making sense of microposts. (#microposts2014) named entity extraction & linking challenge. In Proceedings of the 4th Workshop on Making Sense of Microposts, Co-Located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, 7 April 2014; Volume 1141, pp. 54–60.
10. Han, X.; Sun, L. A generative entity-mention model for linking entities with knowledge base. In Proceedings of the 49th ACL Meeting, Portland, OR, USA, 19–24 June 2011; pp. 945–954.
11. Hachey, B.; Radford, W.; Nothman, J.; Honnibal, M.; Curran, J.R. Evaluating entity linking with Wikipedia. *Artif. Intell.* **2013**, *194*, 130–150. [CrossRef]

12. Moncla, L.; Gaio, M.; Nogueras-Iso, J.; Mustière, S. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *Int. J. Geogr. Inf. Sci. (IJGIS)* **2016**, *30*, 1137–1160. [CrossRef]
13. MacDonald, D. Internal and external evidence in the identification and semantic categorisation of Proper Names. In *Corpus Processing for Lexical Acquisition*, Branimir, B., James, P., Eds.; The MIT Press: Cambridge, MA, USA, 1996.
14. Ait-Mokhtar, S.; Chanod, J. Incremental Finite-State Parsing. In Proceedings of the 5th Applied Natural Language Processing Conference, ANLP 1997, Marriott Hotel, WA, USA, 31 March–3 April 1997; pp. 72–79.
15. Hobbs, J.; Appelt, D.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; Tyson, M. A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In *Finite State Devices for Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1996; pp. 383–406.
16. Friburger, N.; Maurel, D. Finite-state transducer cascade to extract named entities in texts. *Theor. Comput. Sci.* **2004**, *313*, 94–104. [CrossRef]
17. Abney, S. Parsing By Chunks. In *Principle-Based Parsing*; Springer: Berlin/Heidelberg, Germany, 1991; pp. 257–278.
18. Abney, S. Partial Parsing via Finite-State Cascades. In Proceedings of the Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic, 12–23 August 1996; pp. 8–15.
19. Kokkinakis, D.; Kokkinakis, S.J. A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, 8–12 June 1999.
20. Alegria, I.; Aranzabe, M.; Ezeiza, N.; Ezeiza, A.; Urizar, R. Using Finite State Technology in Natural Language Processing of Basque. In *Implementation and Application of Automata*; Watson, B.W., Wood, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 1–12.
21. Makhoul, J.; Kubala, J.; Schwartz, R.; Weischedel, R. Performance measures for information extraction. In Proceedings of the DARPA Broadcast News Workshop, Herndon, VA, USA, 28 February–3 March 1999.
22. Sanabria, C.A.R.D.J. Performance Evaluation of Container-based Virtualization for High Performance Computing Environments. *arXiv* **2017**, arXiv:1709.10140.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).