Developing an Online Indonesian Corpora Repository[*]

Ruli Manurung, Bayu Distiawan, and Desmond Darma Putra

Faculty of Computer Science, Universitas Indonesia,
Depok 16424
maruli@cs.ui.ac.id, bayu.distiawan@ui.ac.id, desmond86@gmail.com

Abstract. This paper describes efforts to develop an online repository of Indonesian corpora –and its associated functions and services– that has been designed to support a wide variety of use cases and applications. Two design considerations are ensuring sustainability and accessibility of the corpora, and enabling open enrichment through annotation. The presented model supports OLAC-compliant metadata, is built atop an OAIS-compliant core repository, and exposes data and functionality via RESTful web services. A prototype implementation is presented, which allows users to upload, browse, and search the collection, whose extensible content model currently supports POS tagging. The future plan is for language-independent aspects of the system to be packaged up and released as an open-source package to aid the development of corpora repositories for other languages.

Keywords: Indonesian, corpora, annotation, metadata, digital repositories.

1 Introduction

Bahasa Indonesia, or just simply Indonesian, is spoken by well over 100 million people, and yet there is a proportionally small amount of available Indonesian language resources that would greatly support linguistics and language technology research. Recent work on Indonesian NLP resources and tools has started to bear results (Adriani and Manurung, 2008), but to further advance research, there is a need for a comprehensive, balanced, and wide-coverage collection of corpora (Arka *et al.*, 2007).

Designing and building an online corpora repository is much more complex than simply uploading a set of text files onto a folder accessible over the Internet. For it to be of support to the research community, careful consideration must be paid to the design of standards, protocols, metadata, and architecture. In this paper we present two main desiderata that inform the design of our corpora repository design: the need to ensure sustainability and accessibility of the corpora (Section 2), and the enabling of open enrichment –through annotation– of the primary data (Section 3), before presenting our design and implemented prototype (Section 4).

2 Ensuring sustainability and accessibility

Simons and Bird (2008) state that for a language corpus to bring benefit, the design must take into consideration the following properties:

- *Extant*. The digital corpus must store data that must be verifiable to be authentic representations of their corresponding artefacts. To maintain this property of data integrity, issues such as backups, authentication, and reliability of storage infrastructure and communication networks must be considered.
- *Discoverable*. An archive that stores any amount of data will only be useful if its contents can be accessed by parties who need the knowledge to accomplish their goals. Thus, the developed

[*] This research is part of a collaborative research project funded by ARC Discovery Grant DP0877595.

archive must be easily discoverable and accessible by stakeholders such as government, researchers, industry, or the general public. Simply making it available over a network is not enough to state that it is discoverable. The data must be easily and efficiently searchable, either through richly annotated metadata, or through intelligent content-based information retrieval systems.

- *Available*. Digital data elements that are extant and discoverable are still not useful if they are not accessible by parties who need them. Thus, data elements in a repository must have clear access policies and clear addressing. On the other hand, making access completely wide-open for valuable date is also undesirable, as there are situations when access must be restricted to certain parties. These issues can be addressed through encryption techniques and role policies. Furthermore, technical issues such as prerequisite bandwidth and computational and storage resources must also be carefully considered, especially given the diverse state of infrastructure throughout Indonesia, which we envisage to be a crucial factor for an online Indonesian corpora repository.
- *Interpretable* & *portable*. Data that is accessible by a user will only be of benefit if it can be displayed and manipulated by appropriate tools. Thus, there must be standardization of formats in order for the data to be independently usable by third parties, without having to consult the producer of the data beforehand. Appropriate standards of protocol and data storage must be determined, e.g. metadata, markup. Additionally, data in the required format should be easily convertible to and from other formats, so that the data can be used in the widest range of scenarios.

To that end, Simons and Bird have established the Open Language Archives Community (OLAC)[1], which specifies a set of standards for metadata, processes, and repositories, which together describes a system by which language resources available on the Internet can be made discoverable through search facilities.

Another related body of work is that on digital preservation, i.e. the management of digital information over time. Textual corpora are not only useful for linguistic research purposes, but also serve as recordings of cultural heritage. Thus, in designing a corpora repository, especially for rare resources of certain languages, one must consider that the digital representation will still be interpretable, say, a hundred years into the future. The Open Archival Information System (OAIS) Reference Model specifies a set of standards and procedures that preserves information in an archive over a long period of time (CCSDS, 2002). In the OAIS model, the environment consists of three participants, i.e. the producer, consumer, and the archive manager, as shown in Figure 1.

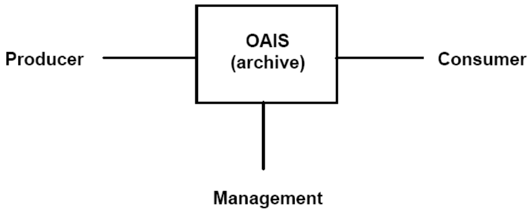


Figure 1: Environment model of the OAIS reference model (CCSDS, 2002).

These three roles must work together so that a piece of information can be interpreted. When a document is *ingested* into the system, it is associated with metadata that forms a Submission Information Package (SIP). In the archive, it may be associated with different metadata that forms an Archival Information Package (AIP), and finally, when the information reaches the consumer, it does so in the form of a Dissemination Information Package (DIP). At each stage, the package contains information on how to interpret the bits that constitute the data.

[1] http://www.language-archives.org

3 Allowing open enrichment of primary data

Corpora can typically be seen as consisting of primary data, i.e. the collected or recorded source material that is the object of observation, and secondary data, i.e. additional information that augments or enhances the knowledge pertaining to the primary data. For linguistic corpora, secondary data takes the form of *annotations* that enrich the linguistic information. These annotations may be produced by the original researchers who first collected the primary data, but may also come from third party researchers if the primary data were made available. For example, a corpus of texts originally collected by a computational linguist, who added POS tags and constituent bracketing, could subsequently be used by a sociolinguist, who then adds annotations concerning register and social context factors.

Thus, one key design consideration is that primary data can be made available for reuse by many communities of researchers. This has implications on the representational design. Annotation data is often stored and manipulated in various formats, thus even if a corpus satisfies the OLAC principles outlined in Section 2, such technical issues can still hamper universal access and sharing of information. There have been a number of efforts to build generalised models of annotation such as annotation graphs (Bird and Liberman, 2001) and the Linguistic Annotation Framework (Ide and Romary, 2007).

Cassidy (2008) presents a web-based interface to annotation data that makes use of an abstract model of annotation, but is able to transform the data into a variety of annotation formats to clients over the web. Key to its approach is (i) the use of current web technologies such as XML and web service standards, in particular representational state transfer (or RESTful) services (Fielding, 2000), to ensure maximum interoperability, and (ii) the adoption of the so-called standoff markup model, which clearly separates primary from secondary data. With standoff markup, annotations (i.e. secondary data) are never embedded ‘inline’ with the primary data, but rather stored in an external location, with pointers that map to specific segments in the primary data. For textual data, the pointers might be to spans of character or tokens, whereas for multimedia data such as audio or video streams, the pointers might be to timestamps that denote signal spans.

4 Corpora repository design

Having considered the issues described in Sections 2 and 3, it is proposed that the Indonesian corpora repository be implemented as a web application with an architecture shown in Figure 2.

There are four main components: (i) the core storage, where the actual primary and secondary

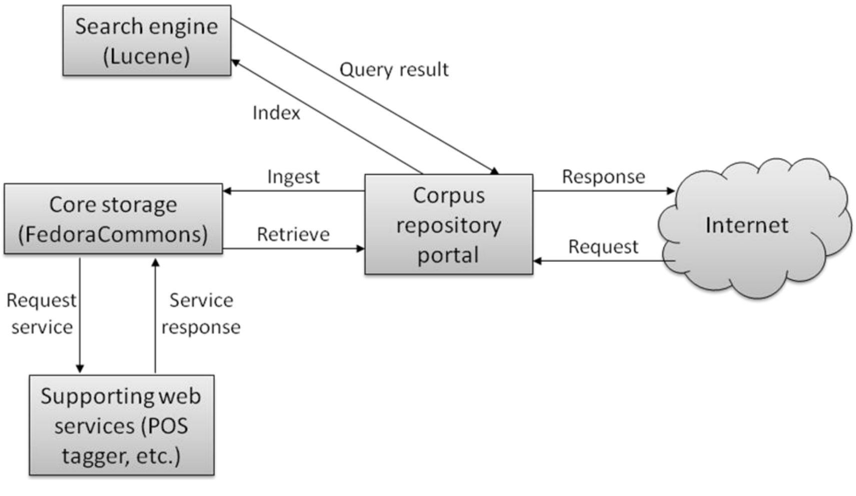


Figure 2: Overview architecture of the proposed corpora repository.

data bits are stored, (ii) the search engine, which maintains content-based indices on the data within the core storage, (iii) a server providing supporting web services as designated by the data content models, and (iv) the corpora repository portal, which serves as the front-end of the system.

4.1 Core storage

The core storage is developed using the Fedora Commons[2] digital object repository, as opposed to a more traditional relational database. Fedora implements the OAIS reference model described in Section 2, and enables a very powerful and extensible platform for defining functionality against the stored data. It stores collections of *digital content objects*, each of which is comprised of certain components, namely: (i) a persistent identifier (PID), which defines a unique address for referencing over the web, (ii) a set of Dublin Core metadata that provides a basic description of the digital object, (iii) one or more *datastreams*, which store the digital information pertaining to the archived object, and (iv) a description of *services* or functions that can be applied to the digital object. The designs of these components for a given object are defined in a *content model*.

In the proposed corpora repository model, the document itself is stored in the main `TextContent` datastream, with subsequent annotations stored in separate datastreams. This affords a standoff markup scheme, so that secondary data is logically separable from primary data, yet remains organised as a single package of information. Moreover, every element of the primary and secondary data can be referenced with a URI, providing a RESTful interface to annotations similar to (Cassidy, 2008). The digital object representation can be seen in Figure 3. We have also defined content models for non-textual data, such as audio and video recordings. Such models will be relevant when considering speech corpora.

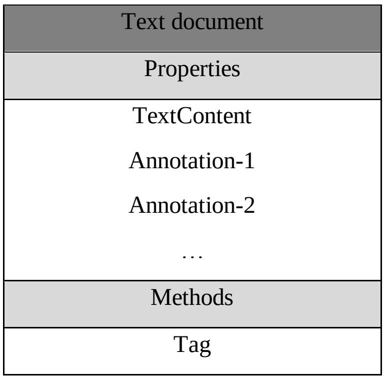


Figure 3: Fedora digital object representation.

In terms of object functionality, the Fedora content model also specifies the services that are available upon the data, through *service definitions* (SDef), which provides an abstract definition of a service that can be applied to an object instance of a particular content model, and *service deployments* (SDep), which represent specific implementations of services as defined by an SDef. Currently, one method has been implemented, **`tag`**, which calls upon a part-of-speech tagger to produce tags for the text within the main datastream. The tagger is implemented as a RESTful web service that wraps an Indonesian POS tagger developed using the Stanford POS tagger (Adriani *et al.*, 2009). Figure 4 displays the service definition of the document content model.

[2] http://www.fedora-commons.org

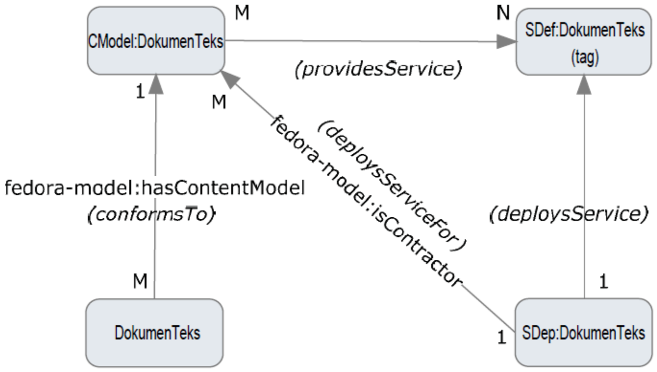


Figure 4: Service definition of the document content model.

4.2 Search engine

Fedora Commons plays the role of a database server in our repository design. However, it does not provide a way of efficient content-based searching of the data[3]. To provide this essential functionality, a search engine component is implemented as a web service external to the core storage, which in turn wraps an indexing machine implemented using the popular open-source Lucene engine[4].

Figure 2 illustrates that the search engine is called by the corpora repository portal. It does not actually store any of the primary or secondary data itself, and instead stores indices that point to the URIs into the core storage. To maintain synchronization between the data stored within the core storage and the search engine index, it is the responsibility of the portal to update the search engine index whenever a document is ingested into the core storage. Fortunately, Lucene provides functionality to do incremental updates to the index, preventing the need to recompute the entire index every time a document is added. The search engine stores a link to the document URI along with a short snippet of the context wherein the query keywords appear.

4.3 Supporting web services

To facilitate interpretation and manipulation of the information stored within the repository, various services can be defined through the content models of the digital objects. As mentioned in Section 4.1, a **`tag`** service has been defined, and is implemented as a RESTful web service that wraps an Indonesian POS tagger developed using the Stanford POS tagger (Adriani *et al.*, 2009).

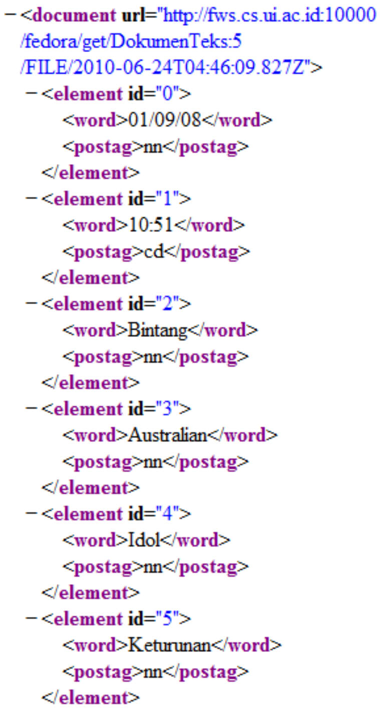
The text in the main datastream is fed to the tagger, which returns an XML document containing a list of enumerated elements, each of which represents a token and its corresponding POS tag. Figure 5 shows a sample output from the tagger web service.

Note that the XML document returned by the POS tagger potentially enables per-token RESTful URI references to the corpus (Cassidy, 2008).

We have also implemented a web service that automatically generates OLAC-compliant metadata based on documents ingested into the repository. Experiments were done using the ILPS collection (Tala, 2003), and were validated to be correct by the OLAC conformance review service.

[3] Although it does provide built in services to query the Dublin Core metadata.

[4] http://lucene.apache.org

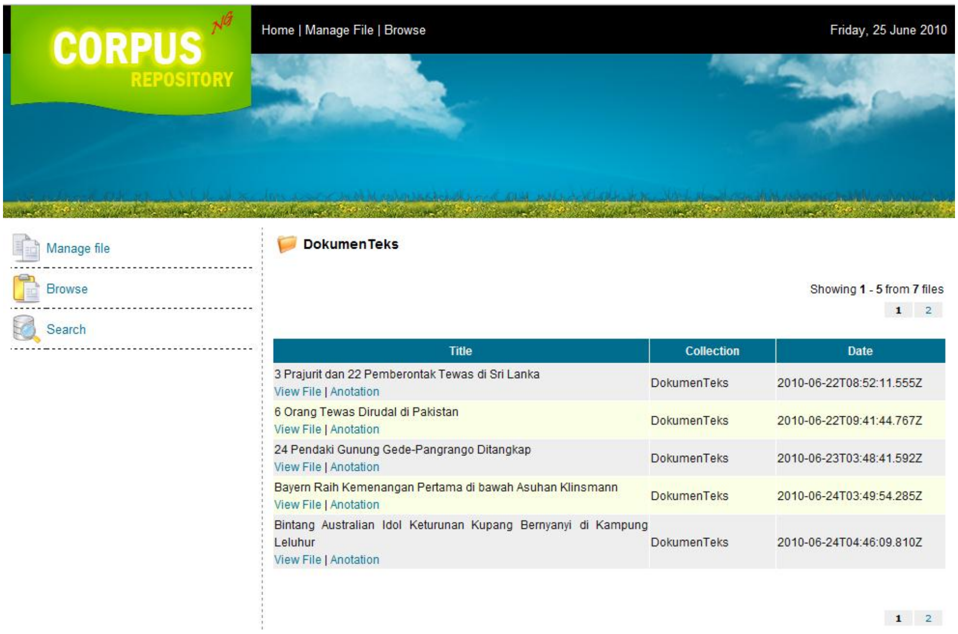


```
- <document url="http://fws.cs.ui.ac.id:10000
  /fedora/get/DokumenTeks:5
  /FILE/2010-06-24T04:46:09.827Z">
  - <element id="0">
      <word>01/09/08</word>
      <postag>nn</postag>
    </element>
  - <element id="1">
      <word>10:51</word>
      <postag>cd</postag>
    </element>
  - <element id="2">
      <word>Bintang</word>
      <postag>nn</postag>
    </element>
  - <element id="3">
      <word>Australian</word>
      <postag>nn</postag>
    </element>
  - <element id="4">
      <word>Idol</word>
      <postag>nn</postag>
    </element>
  - <element id="5">
      <word>Keturunan</word>
      <postag>nn</postag>
    </element>
```

Figure 5: Sample XML output of the POS tagger web service

4.4 Corpora repository portal

Finally, the corpora repository itself is accessible through a web portal, which provides the functionality to browse, upload, and search through the corpora. A current prototype has been deployed at http://bahasa.cs.ui.ac.id/corpusRepository. It also provides access to all services defined on the document content model, e.g. the POS tagging service. A screenshot of the portal can be seen in Figure 6.



DokumenTeks

Showing **1** - **5** from **7** files

Title	Collection	Date
3 Prajurit dan 22 Pemberontak Tewas di Sri Lanka View File | Anotation	DokumenTeks	2010-06-22T08:52:11.555Z
6 Orang Tewas Dirudal di Pakistan View File | Anotation	DokumenTeks	2010-06-22T09:41:44.767Z
24 Pendaki Gunung Gede-Pangrango Ditangkap View File | Anotation	DokumenTeks	2010-06-23T03:48:41.592Z
Bayern Raih Kemenangan Pertama di bawah Asuhan Klinsmann View File | Anotation	DokumenTeks	2010-06-24T03:49:54.285Z
Bintang Australian Idol Keturunan Kupang Bernyanyi di Kampung Leluhur View File | Anotation	DokumenTeks	2010-06-24T04:46:09.810Z

Figure 6: Screenshot of the prototype implementation.

5 Summary

We have presented the design considerations for a model and prototype implementation of an online corpora repository that has been carefully considered to support a wide variety of use cases over long periods of time. It supports OLAC-compliant metadata, is built atop an OAIS-compliant repository system (Fedora Commons), and provides RESTful web service interfaces for accessing data and functions on the data. The initial prototype is available at http://bahasa.cs.ui.ac.id/corpusRepository, and we intend to continue development, and to populate the repository with various corpus collections. Additionally, the repository is already equipped to ingest non-textual data such as audio and video recordings. It is hoped that this will aid further development into language technology research on the Indonesian language. Planned future functionality includes the ability for researchers to add their own annotations to existing corpora, directly from the portal website. This functionality could utilize rich interactive AJAX technology, e.g. the Serengeti annotator[5]. The future plan is for language-independent aspects of the system to be packaged up and released as an open-source package to aid the development of corpora repositories for other languages.

References

Adriani, M. and R. Manurung. 2008. A Survey of Bahasa Indonesia NLP Research Conducted at the University of Indonesia. *Proceedings of the 2nd International MALINDO Workshop*. Cyberjaya, Malaysia.

Adriani, M., R. Manurung, and F. Pisceldo. 2009. Statistical Based Part Of Speech Tagger for Bahasa Indonesia. *Proceedings of the 3rd International MALINDO Workshop*. ACL-IJCNLP 2009. Singapore.

Arka, I. W., J. Simpson, A. Andrews, and M. Dalrymple. 2007. Challenges of Developing a Balanced and Representative Corpus for Indonesian ParGram. *Proceedings of the 11th Internatonal Symposium on Malay/Indonesian Linguistics*, Manokwari, Indonesia.

Bird, S. and M. Liberman. 2001. A Formal Framework for Linguistics Annotation. *Speech Communication*, 33:1-2, pp.23-60.

Cassidy, S. 2008. A RESTful Interface to Annotations on the Web. *Proceedings of the 2nd Linguistic Annotation Workshop (LAW II)*, LREC2008, Marrakech.

CCSDS. 2002. Reference model for an Open Archival Information System (OAIS). Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems. Also published as ISO 14721:2003.

Fielding, R. T. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine.

Ide, N. and L. Romary. 2007. Towards International Standards for Language Resources. In L. Dybkjaer, H. Hemsen, and W. Minker, editors, *Evaluation of Text and Speech Systems*, pages 263–84. Springer.

Simons, G. and S. Bird. 2008. Toward a Global Infrastructure for the Sustainability of Language Resources. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pp.87-100, Cebu, Philippines.

Tala, F. Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. M.Sc. Thesis, University of Amsterdam.

[5] http://coli.lili.uni-bielefeld.de/serengeti