

# Glottocodes: Identifiers Linking Families, Languages and Dialects

Robert Forkel<sup>a,\*</sup> and Harald Hammarström<sup>b</sup>

<sup>a</sup> *Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Germany*

*E-mail: forkel@shh.mpg.de*

<sup>b</sup> *Department of Linguistics and Philology, Uppsala University, Sweden*

*E-mail: harald.hammarstrom@lingfil.uu.se*

**Abstract.** Glottocodes constitute the backbone identification system for the language, dialect and family inventory Glottolog (<https://glottolog.org>). In this paper, we summarize the motivation and history behind the system of glottocodes and describe the principles and practices of data curation, technical infrastructure and update/version-tracking systematics. Since our understanding of the target domain — the dialects, languages and language families of the entire world — is continually evolving, changes and updates are relatively common. The resulting data is assessed in terms of the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship. As such the glottocode-system responds to an important challenge in the realm of Linguistic Linked Data with numerous NLP applications.

**Keywords:** Linguistics, Linked Data, Language Inventory, Linguistic Standards

## 1. Introduction

Glottocodes constitute the backbone identification system for the language, dialect and family inventory Glottolog (<https://glottolog.org>, currently in edition 4.3, [1]). A glottocode consists of four alphanumeric characters (i.e., lowercase letters or decimal digits) and four decimal digits, for example `abcd1234` or `b10b1234`. Glottocodes are complementary to three-letter ISO 639-3 language identification codes (see <https://iso639-3.sil.org/>) which, however, concern languages only.

In the current release, there are over 25 695 glottocodes (8 525 language-level, 4 544 family-level and 12 626 dialect-level).

## 2. Motivation and History

Glottocodes were introduced in 2010 by Glottolog collaborator Sebastian Nordhoff, in response to the following requirements:

- An ID specifically devoted to machine treatment, not confusable with an informal or human-directed identifier
- An ID type oblivious to level of linguistic abstraction (idiolect, sociolect, dialect, language, sub-family, family, etc.)
- An ID system for languages that improves on the ISO 639-3 language identifiers in terms of quality, transparency and anchoring

The 8-character long alphanumeric string was designed not to resemble an abbreviation or an easily remembered mnemonic. This was done specifically in order to counter any temptation to capitalize, modify, inflect or translate it, which users might if the ID-string had had a more human-palatable appearance (such as a three-letter mnemonic, a standardized name or the like).

---

\*Corresponding author. E-mail: forkel@shh.mpg.de.

1 Glottolog had adopted a doculect-based approach<sup>1</sup>  
 2 for organizing concrete attestations of languages as  
 3 recorded by bibliographical references. This means  
 4 that language data (ultimately emanating from idi-  
 5 olects of specific speakers) recorded in different publi-  
 6 cations are grouped into successively larger conglom-  
 7 erates such as subdialects, dialects, languages, subfam-  
 8 ilies and families [7]. Earlier approaches had always  
 9 sought to tie the identity of a set of data to a specific  
 10 level such as language or dialect, such that if the set  
 11 of data did not change but the level changed (e.g., a  
 12 language reconsidered a dialect or vice versa) the ID  
 13 had to change as well (e.g., [8]). Given controversies  
 14 over language and dialect status, along with our in-  
 15 complete understanding of the language situation for  
 16 many minority languages around the world, such level  
 17 changes are actually not uncommon. With the level-  
 18 independent ID-system of glottocodes, merely a level-  
 19 attribute, but not the ID itself would have to change in  
 20 such cases. The level-neutral term for the denotation of  
 21 a glottocode is *languoid* [2].

22 Finally, a decade or longer ago, the quality and  
 23 transparency of the ISO 639-3 standard for languages  
 24 was problematic and an alternative was clearly needed  
 25 [9]. Fortunately, the quality and transparency has im-  
 26 proved in the meantime so that the discrepancy be-  
 27 tween language-level glottocodes and ISO 639-3 codes  
 28 is diminishing (but not completely eradicated, see 3.3).

### 32 3. Glottolog data curation

33 As explained above, glottocodes are the identifiers  
 34 for *languoids* — the main objects of the data curated  
 35 in Glottolog.

36 In the following, we briefly describe the infrastruc-  
 37 ture and framework pioneered for the Glottolog data  
 38 for data curation and publication ([10] and [11]).

- 39 – All data is stored in UTF-8 encoded text files  
 40 (with consistency ensured by the `pyglottolog`  
 41 software package ([12])).
- 42 – Thus, collaboration and curation workflows can  
 43 make use of `git` a distributed version control  
 44 system to track history of changes and prove-  
 45 nance.

46  
 47  
 48  
 49  
 50 <sup>1</sup>The term 'doculect' emanates from [2] but the general approach  
 51 has been used by many earlier authors, notably [3–6].

- 1 – The master copy of the Glottolog `git` reposi-  
 2 tory is hosted with GitHub at [https://github.com/  
 3 glottolog/glottolog](https://github.com/glottolog/glottolog) for curation (but due to the  
 4 nature of `git` repositories — where all essential  
 5 metadata is part of the every copy of the reposi-  
 6 tory — this does not establish lock-in).
- 7 – Released versions of the repository are published  
 8 and archived with Zenodo (<https://zenodo.org>).

9 This setup does not only provide a stable manage-  
 10 ment system for all information shared between the  
 11 Glottolog editors, but also a collaborative environment  
 12 which allows involving the wider community. Some-  
 13 what similar to — but less formal than — ISO 639-  
 14 3 change requests, Glottolog users can make use of  
 15 GitHub issues to indicate errors or request inclusion of  
 16 new languoids<sup>2</sup>, or even submit pull requests with data  
 17 corrections<sup>3</sup>.

18 Glottolog aims to share this data in an open and  
 19 FAIR ([13]) way. Stepping through the FAIR Guiding  
 20 Principles for scientific data management and steward-  
 21 ship<sup>4</sup> will shed light on the details.

#### 22 3.1. Glottolog data is findable

23 Glottolog is a well established language catalog  
 24 as evidenced by more than 400 citations of editions  
 25 of Glottolog such as "Glottolog 4.0" in the scholarly  
 26 literature (according to Google Scholar). The Glot-  
 27 tolog data repository lists 16 contributors in addition  
 28 to the Glottolog editors (and not including users open-  
 29 ing issues) (see [https://github.com/glottolog/glottolog/  
 30 graphs/contributors](https://github.com/glottolog/glottolog/graphs/contributors)) — pointing to a healthy, collabora-  
 31 tive user community.

32 *Glottolog data are registered and indexed in a  
 33 searchable resource.*

34 Glottolog data is also well indexed in relevant cata-  
 35 logues: The first point of contact for many users is the  
 36 Glottolog web application at <https://glottolog.org> —  
 37 not at least because it is well indexed by Google and  
 38 other search engines. But Glottolog data is also har-  
 39 vested by OLAC (Open Language Archives Commu-  
 40 nity) and is listed in the OLAC catalogue as the archive  
 41 with the highest number of distinct languages (see  
 42 <http://www.language-archives.org/metrics/glottolog.org>).  
 43 Finally, Glottolog data releases can be found on Zen-

44  
 45  
 46  
 47  
 48  
 49 <sup>2</sup>E.g. <https://github.com/glottolog/glottolog/issues/646>

<sup>3</sup>E.g. <https://github.com/glottolog/glottolog/pull/648>

<sup>4</sup><https://www.go-fair.org/fair-principles/>

odo, and consequently wherever Zenodo metadata is indexed.

*Glottolog data are assigned a globally unique and persistent identifier.*

All languoids in Glottolog are unambiguously identified via glottocodes. These glottocodes are transparently associated with URLs in the glottolog.org domain, turning them into globally unique identifiers. Each release of Glottolog is identified by the DOI assigned by Zenodo.

*Metadata clearly and explicitly include the identifier of the data they describe.*

CLDF — one of the dissemination formats of Glottolog — is designed to allow for explicit linking of metadata to identifiers. The underlying mechanism to do this is described in [14], and the semantics are provided through the CLDF Ontology (see <https://cldf.cldf.org/v1.0/terms.rdf>).

### 3.2. Glottolog data is accessible

*Glottolog data are retrievable by their identifier using a standardised communications protocol.*

Zenodo (and the metadata associated by Zenodo with the DOI assigned to data releases) guarantees that data is retrievable using the standard protocol associated with DOIs.

For each languoid, the CLDF/CSVW data associates an HTTP([15]) URL, which is resolvable via the Glottolog web application.

### 3.3. Glottolog data is interoperable

*Glottolog data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

Glottolog aims at integration with the Semantic Web at large and the Linguistic Linked Data initiative in particular.

At the most fundamental level this means resource URLs — aka URLs for Glottocodes. These resource URLs are not only usable as universally unique identifiers, but are also resolvable through the Glottolog web application. HTTP status codes ([15]) returned by the web application signal the status of Glottocodes as follows:

**200 OK** for active codes

**410 GONE** or 301 MOVED PERMANENTLY for retired Glottocodes

**404 NOT FOUND** for invalid codes

The web application also provides several serializations of RDF ([16]) representations of languoid data. These serializations can be retrieved using standard content negotiation mechanisms such as using ACCEPT HTTP headers.

While these efforts provide convenient integration with the "living" Semantic Web, Glottolog also aims at interoperability for its archived, long-term available datasets. To this end, Glottolog data is serialized as CLDF Structure Dataset ([17]). The CLDF standard ([18], [19]) does not only provide interoperability with other CLDF datasets, but — due to being built on the W3C's "CSV on the Web" recommendation ([14] and [20]) — also allows automatic conversion to RDF ([21]).

*Glottolog data use vocabularies that follow FAIR principles.*

CLDF bundles data with structured, machine readable, semantic web-ready metadata. Since CLDF metadata is encoded in JSON-LD ([22]), the data can be marked up using standard ontologies such as Dublin Core, DCAT (<https://www.w3.org/ns/dcat#>) and PROV (<https://www.w3.org/ns/prov#>).

*Glottolog data include qualified references to other (meta)data.*

Thanks to last-decade improvements in the curation of ISO 639-3 language identifiers, ISO 639-3 codes and language-level glottocodes are one-to-one interchangeable for the vast majority of cases, and the differences are few enough that a specific comment explaining the differences are given in each of the remaining cases on Glottolog. In fact, Glottolog aims at covering all valid ISO 639-3 codes to provide a full mapping, but typically there is a time lag of a couple of months between additions to ISO 639-3 and a Glottolog release addressing these changes. There remains a principled difference in anchoring where the denotation of a glottocode in Glottolog is defined by the data and information in the references tied to it. The references are associated in Glottolog in such a way that the referenced data and information is enough to distinguish the languoid from all other languoids. In ISO 639-3, each entry has metadata such as geographical information, name(s), speaker numbers and classification which presumably defines the language,

1 but no actual or referenced data from the denoted language.  
2 Unfortunately, it is not so that metadata information is in all cases enough to identify its denotation.  
3 Language names are notoriously ambiguous and  
4 the case of language-shifting ethnic groups is particularly tricky, as most metadata (speaker numbers, geography, name) is not sufficient to disambiguate between  
5 the original and substituted language.  
6

7  
8  
9 Glottolog, and in particular individual languoids, are also well-linked from Wikipedia. In particular,  
10 practically all language- and family- level are referenced in Wikipedia. These Wikipedia links translate  
11 to Wikidata links (e.g. <https://www.wikidata.org/wiki/Q31746>) which in turn provide links to other language  
12 identification schemes such as ISO 639-2 (which are arguably less important than Ethnologue/ISO 639-3 in  
13 the contexts where Glottolog is most used).  
14  
15  
16  
17  
18

### 19 3.4. *Glottolog data is reusable*

20  
21 *Glottolog data are released with a clear and accessible data usage license.*  
22

23 Glottolog data is release under a Creative Commons  
24 CC-BY-4.0 license.

25 *Glottolog data are associated with detailed provenance.*  
26  
27

28 Like most large-scale databases, parts of Glottolog  
29 data are aggregated from various sources. Glottolog  
30 tries to be transparent about this, e.g. by  
31

- 32 – providing references for all classification proposals<sup>5</sup>
- 33 – providing references for all endangerment assessments<sup>6</sup>
- 34 – describing the provenance of the bibliography<sup>7</sup>

35  
36  
37 *Glottolog data meet domain-relevant community standards.*  
38  
39

40 We already described the relation between glottocodes and ISO 639-3 language codes. Arguably, the  
41 transparent mapping between the two, which Glottolog provides, is the most important contribution towards  
42 meeting domain-relevant standards.  
43  
44

45 But as explained above, Glottolog also  
46

47 <sup>5</sup>E.g. <https://github.com/glottolog/glottolog/blob/v4.3/languoids/tree/indo1319/anat1257/luvi1234/cari1274/md.ini#L59-L63>

48 <sup>6</sup>Cf. [https://github.com/glottolog/glottolog/blob/v4.3/config/aes\\_sources.ini](https://github.com/glottolog/glottolog/blob/v4.3/config/aes_sources.ini)

49 <sup>7</sup>Cf. <https://github.com/glottolog/glottolog/blob/v4.3/references/BIBFILES.ini>  
50  
51

- 1 – caters to the LLD community, by meeting Semantic Web standards,
- 2
- 3 – serves the OLAC community, by implementing the OAI-PMH data provider specification ([23]),
- 4 thereby allowing harvesting through OLAC,
- 5
- 6 – helps researchers in Diversity Linguistics to inform their analyses using Glottolog metadata, by
- 7 making this data accessible as CLDF dataset.  
8

## 9 4. Policies governing glottocode assignment

10  
11  
12  
13 Glottolog aims to be complete with respect to all assertable L1 languages<sup>8</sup> in the real world, so all languages in the world (as far as this is understood at the  
14 time of a certain release) have a language-level glottocode. Glottolog makes a classification decision for  
15 all and only language-level languoids so the family-level inventory is complete in the sense of exhausting  
16 the languages of a given release.  
17  
18

19 In contrast, the inventory for dialects (variety of a language), non-L1 languages (artificial languages, speech registers, pidgins) and non-assertable languages (putative languages for which there is insufficient data to decide if they are different from all other languages) and putative families (hypotheses about family relationships that have appeared in the literature) is ambitious but not complete. The world may contain more of any or all of these entities without a necessary reflection in a glottocode. Genuine completeness with respect to these categories is deemed practically (if not theoretically) impossible.  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32

33 For these reasons, Glottolog accounts for any changes to the language-level inventory between two releases, so if something was deemed a real-world language, a user can follow any changes to that assertion. If a language-level languoid was completely erroneous, it is moved to the Bookkeeping category. If it is promoted/demoted to a family/dialect, it retains its glottocode but changes its level accordingly, but from that point on it ceases to be “protected” by its language-level status. In contrast, the family-level and dialect-level glottocodes are not “protected” and may be removed from the inventory between releases. Since they do not necessarily reflect a real-world entity like an L1 language, it cannot systematically be explained what “happened” to them, e.g., if they never really existed. However, some tracking possibilities are always guar-  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

50 <sup>8</sup>See <https://glottolog.org/glottolog/glottologinformation> for an explanation of these criteria.  
51

anted because both families and dialects are linked to language-level languoids. If a family-level glottocode disappears, it is possible to check which language-level languoids it covered and to check which family/ies they are now associated with and if a dialect-level glottocode disappears, it is possible to check which language-level languoid it pertained to and to check which dialects are now associated with it. Furthermore, the git-versioning and the structure of the index allows a quick location of the specific pull-request associated with the removal/appearance of a glottocode (see 5).

Glottocodes are not recycled — for new entities, completely new glottocodes are assigned (retired codes are not re-used/re-purposed). Hence, all glottocodes that have ever appeared are either active or retired.

## 5. Glottolog versioning

We already pointed out that Glottolog data is versioned and released periodically (aiming at a bi-annual release frequency). Each such Glottolog release is self-contained, i.e. does not reference any base data, but instead includes it. Thus, when linking other resources to Glottolog, one should always specify the particular target version.

Glottolog follows a semantic-versioning scheme<sup>9</sup> for data:

- Resources using Glottolog should always target the highest patch version of a particular minor version. This should not break any processing code, but may correct errata.
- Upgrading resources to a new minor version may change data/links, but should not break processing code.
- Upgrading to a new major version may break processing code - i.e. the data structure may change.

The Glottolog version history can be explored in two ways: The Glottolog web application resolves resource URLs of obsolete languoids as follows

```
$ curl -I https://glottolog.org/resource/languoid/id/awun1244
HTTP/1.1 301 Moved Permanently
Date: Fri, 08 Jan 2021 12:06:32 GMT
Content-Type: text/html; charset=UTF-8
Location: https://glottolog.org/files/glottolog-4.0/awun1244.html
```

<sup>9</sup>Cf. <https://semver.org/>

Where the HTML page at <https://glottolog.org/files/glottolog-4.0/awun1244.html> provides context about the languoid and versions when it was still active. In the case of obsolete dialect-level languoids — such as Awuna — this typically allows determining the parent language-level languoid, which will always be part of the current release.

Alternatively, since Glottolog data is curated as git repository, we can use the git software to inspect the history. In the case of Awuna (awun1244), e.g., we learn that it was removed during a "cleanup" of the classification of the Gbe sub-family:

```
$ git log --all --full-history -- "**/awun1244/md.ini"
commit 923cd9eb21f27bc9ae0797a315dd48fd009e53c4
Author: d97hah <harald@bombo.se>
Date: Thu Jul 25 14:50:11 2019 +0200
    Gbe (#387)
    * Gbe resolved clf + move of some Mixed languages
    * fix to phoenician-punic-ugaritic
    * New hh.bib + clf ref updates
    * synched refs
```

## 6. Conclusion

We have described the practices and principles for glottocodes as the identificational system for the languages, dialects and families of the world including data curation, technical infrastructure and update/version-tracking systematics. The resulting data observes the crucial aspects of the FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship. As such the glottocode-system responds to an important challenge in the realm of Linguistic Linked Data with numerous NLP applications.

## References

- [1] H. Hammarström, R. Forkel, M. Haspelmath and S. Bank, *glottolog/glottolog: Glottolog database 4.3*, Zenodo, 2020. doi:10.5281/zenodo.4061162.
- [2] M. Cysouw and J. Good, Languoid, Doculect, Glossonym: Formalizing the notion "language", *Language Documentation and Conservation* 7 (2013), 331–359.
- [3] W. Schmidt, Gliederung der australischen Sprachen, *Anthropos* 7, 7, 7, 8, 9, 12/13, 12/13 (1912, 1912, 1912, 1913, 1914, 1917/1918, 1917/1918), 230–251, 463–497, 1014–1048, 526–554, 980–1018, 437–493, 747–817.
- [4] Č. Loukotka, *Classification of the South American Indian Languages*, Reference Series, Vol. 7, Los Angeles: Latin American Center, University of California, 1968.
- [5] G. van Bulck, *Les recherches linguistiques au Congo Belge: résultats acquis, nouvelles enquêtes à entreprendre*, Mémoires de l'Institut Royal Colonial Belge, Bruxelles, Vol. 16, Bruxelles: Institut Royal Colonial Belge, Bruxelles, 1948.

- [6] J. Good and C. Hendryx-Parker, Modeling Contested Categorization in Linguistic Databases, in: *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, Michigan. June 20-22, 2006*, Lansing, Michigan: E-MELD, 2006, pp. 1–22.
- [7] S. Nordhoff and H. Hammarström, Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources, in: *Proceedings of the First International Workshop on Linked Science 2011*, T. Kauppinen, L.C. Pouchard and C. KeBler, eds, CEUR Workshop Proceedings, Vol. 783, CEUR, 2011, pp. 1–7. <http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/LISC/nordhoff.pdf>.
- [8] M. Mann and D. Dalby, *A thesaurus of African languages: a classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation*, London: Hans Zell, London, 1987. ISBN 9780905450247.
- [9] H. Hammarström, *Ethnologue* 16/17/18th editions: A comprehensive review, *Language* 91(3) (2015), 723–737, Plus 188pp online appendix..
- [10] R. Forkel, Glottolog 3.0 released, 2017. <https://clld.org/2017/03/29/glottolog-3-0.html>.
- [11] R. Forkel, Glottolog 3.0 – A collaborative, versioned catalog of languages and dialects, 2016. <https://clld.org/docs/poznan/glottolog-3-0.pdf>.
- [12] R. Forkel, S.J. Greenhill and C. Rzymiski, glottolog/pyglottolog: Glottolog API, Zenodo, 2020. doi:10.5281/zenodo.3753876.
- [13] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A. 't Hoen, R. Hoofit, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3(160018) (2016), 1–9.
- [14] J. Tennison, G. Kellogg and I. Herman, Model for Tabular Data and Metadata on the Web, Technical Report, World Wide Web Consortium (W3C), 2015. <http://www.w3.org/TR/tabular-data-model/>.
- [15] R. Fielding and J. Reschke, Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, RFC, 7231, RFC Editor, 2014. ISSN 2070-1721. <https://www.rfc-editor.org/rfc/rfc7231.txt>.
- [16] F. Manola and E. Miller, RDF Primer, Technical Report, W3C, 2004. <https://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
- [17] H. Hammarström, R. Forkel, M. Haspelmath and S. Bank, glottolog/glottolog-cldf: Glottolog database 4.3 as CLDF, Zenodo, 2020. doi:10.5281/zenodo.4061165.
- [18] R. Forkel, J.-M. List, S.J. Greenhill, C. Rzymiski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G.A. Kaiping and R.D. Gray, Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics, *Scientific Data* 5(180205) (2018), 1–10. doi:<https://doi.org/10.1038/sdata.2018.205>. <https://www.nature.com/articles/sdata2018205>.
- [19] R. Forkel, J.-M. List, M. Cysouw and S.J. Greenhill, CLDF 1.0, Technical Report, Max Planck Institute for the Science of Human History, Jena, 2017. doi:10.5281/zenodo.1117644.
- [20] R. Pollock, J. Tennison, G. Kellogg and I. Herman, Metadata Vocabulary for Tabular Data, Technical Report, World Wide Web Consortium (W3C), 2015. <https://www.w3.org/TR/tabular-metadata/>.
- [21] J. Tandy, I. Herman and G. Kellogg, Generating RDF from Tabular Data on the Web, Technical Report, World Wide Web Consortium (W3C), 2015. <https://www.w3.org/TR/csv2rdf/>.
- [22] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin and N. Lindström, JSON-LD 1.1 – A JSON-based Serialization for Linked Data, Technical Report, World Wide Web Consortium (W3C), 2020. <https://www.w3.org/TR/json-ld/>.
- [23] C. Lagoze, H.V. e Sompel, M. Nelson and S. Warner, The Open Archives Initiative Protocol for Metadata Harvesting, Technical Report, Open Archives Initiative, 2002. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.