Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users

Ulrich Apel
Eberhard Karls University Tübingen
ulrich.apel@uni-tuebingen.de

Abstract

The project *WaDokuJT* (2013) has developed into the most comprehensive Japanese-German dictionary in regard to covered lemmata and translation equivalents. The dictionary is used by many users from German speaking countries and from Japan. The project was designed concentrating on German speaking users who want to read and translate Japanese texts. So, priority is given to a rather complete coverage of orthographic variations of Japanese headwords on the one side, and German translation equivalents on the other side. Definitions are given only when they are really necessary and only in German language; further, they are as short as possible.

This means that the dictionary project may have certain shortcomings for text production, for Japanese users or for users who need deeper encyclopedic explanations connected to a certain dictionary headword.

This paper presents an approach how to alleviate such deficiencies by providing hyperlinks or other references to data of other dictionaries and encyclopedia projects which may contain the information, users are looking for. For this aim, *WaDokuJT* data refers especially to open source projects, dictionaries, which aren't protected by copyright anymore, and collaborating projects, the data of which can be accessed easily at least in parts.

Keywords: References; Japanese; German; hyperlinks

1 *WaDokuJT* – Project overview

The Japanese-German dictionary project *WaDokuJT* (2013) started in 1998 as individual initiative at Osaka University (Apel, 2001). In the meantime it covers approximately 115,000 lemmata and around 275,000 database records. Records like derivations, compounds, examples of usage and example sentences are not main headwords.

Records contain the Japanese lemmata, their pronunciation written in Japanese syllable alphabet – *hiragana* – and translation equivalents. Orthographical variety in Japanese is represented in more than half a million different written forms of the entries. The *kana* transliteration is extended by a mark-up to calculate Rōmaji transcription following the new standard *DIN 32708 – Transliteration of*

Japanese (*NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen*, 2013). The number of translation equivalents is more than half a million.

2 Weakness investigation of the project and possible improvements

The project was designed mainly for text reception of German users. Sometimes additional information like domain, definitions, explanations etc. are given in German, too. Some information for text production is added, like the Japanese pitch accent for correct pronunciation or conjugation types for Japanese verbs. A certain concession for Japanese users is the addition of a mark-up for the gender of German headwords within the German translation equivalents.

The dictionary's limitations seem to be not too serious, as the dictionary is used by Japanese users and also for text production.[1] Nevertheless, improvements are possible. Further information could be added directly to the dictionary with a specific mark-up and would be displayed only for certain user profiles or use situations. Unfortunately this would complicate the data management not to mention the lack of resources to provide such detailed information. Another way to make further information available without too much trouble for the users and lexicographers is to include links and references to other projects or dictionaries with additional information.

Giving the example of this dictionary project, the paper presents how such links can be added by hand or with automated suggestions of link candidates.

3 Linking to data in public domain or with open source license

3.1 Großes Japanisch-Deutsches Wörterbuch from 1937

The most comprehensive printed Japanese-German dictionary was edited by Kinji Kimura and was published first in 1937. It is still reprinted since then (*Großes Japanisch-Deutsches Wörterbuch*, 1952). The

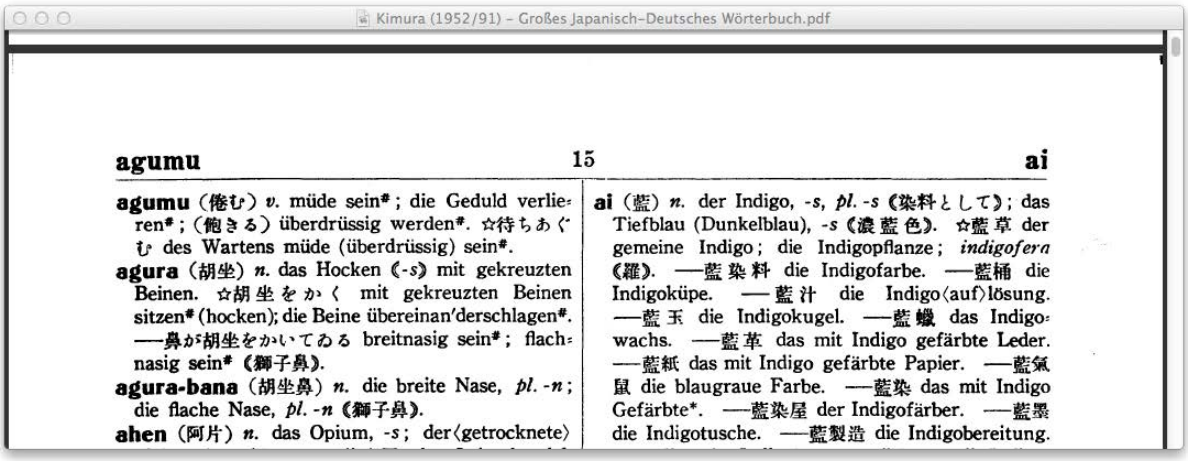
1 Actually, the dictionary is often used as German-Japanese dictionary, since the German translation side can be easily searched, too. This of course is not recommended because users may get overwhelmed by possible Japanese translations, some of which are outdated or are used only in rather specific situations. Since there is quite a number of rather good German-Japanese dictionaries, there would be enough alternatives, although most of these dictionaries concentrate on the needs of text reception by Japanese users (e.g. *Shogakukan Großes Deutsch-Japanisches Wörterbuch*, 1998, or *Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch*, 1989).

author deceased in 1948, and the copyright expired in 1998 – fifty years later according to Japanese copyright laws.

This dictionary seems to be written with the purpose to enable Japanese diplomats and writers to explain Japanese intentions and ways of thinking to Westerners and especially to speakers of German. Different meanings of polysemous words are defined in Japanese and the corresponding translation equivalents are given in German. The German equivalents are completed by articles to show gender, by verb conjugation, by noun and adjective declension or by adjective comparison etc.

This information is still rather valuable, especially for Japanese users and for German text production, even though the Kimura dictionary is outdated in certain aspects. For example, it doesn't reflect orthographic reforms in Japan and Germany or it contains many Japanese names for cities in Manchuria, which was *de facto* a Japanese colony at the time of the compilation. Nowadays, these names are of very limited use.

For German users the Kimura dictionary is also lacking for example the pronunciation of Japanese subentries, the conjugation of Japanese verbs and further German explanations.


Kimura (1952/91) – Großes Japanisch-Deutsches Wörterbuch.pdf

agumu 15 ai

agumu (倦む) *v.* müde sein#; die Geduld verlieren#; (飽きる) überdrüssig werden#. ☆待ちあぐむ des Wartens müde (überdrüssig) sein#.
agura (胡坐) *n.* das Hocken (*-s*) mit gekreuzten Beinen. ☆胡坐をかく mit gekreuzten Beinen sitzen# (hocken); die Beine übereinan'derschlagen#. ——鼻が胡坐をかいてゐる breitnasig sein#; flachnasig sein# (獅子鼻).
agura-bana (胡坐鼻) *n.* die breite Nase, *pl.* *-n*; die flache Nase, *pl.* *-n* (獅子鼻).
ahen (阿片) *n.* das Opium, *-s*; der⟨getrocknete⟩

ai (藍) *n.* der Indigo, *-s*, *pl.* *-s* (染料として); das Tiefblau (Dunkelblau), *-s* (濃藍色). ☆藍草 der gemeine Indigo; die Indigopflanze; *indigofera* (羅). ——藍染料 die Indigofarbe. ——藍桶 die Indigoküpe. ——藍汁 die Indigo⟨auf⟩lösung. ——藍玉 die Indigokugel. ——藍蠟 das Indigowachs. ——藍革 das mit Indigo gefärbte Leder. ——藍紙 das mit Indigo gefärbte Papier. ——藍氣鼠 die blaugraue Farbe. ——藍染 das mit Indigo Gefärbte*. ——藍染屋 der Indigofärber. ——藍墨 die Indigotusche. ——藍製造 die Indigobereitung.

Figure 1: Screenshot of a sample page of the *Großes Japanisch-Deutsches Wörterbuch* (1952) in a digitalized version as PDF with the example entry of *agura*.

The *WaDokuJT* project is now adding gradually tags that refer from one *WaDokuJT* entry to the corresponding page, column and entry count of the Kimura dictionary. This process involves quite a lot of manual operations, but it can be supported by electronic means. A scan of the Kimura dictionary had enough quality to run an optical character recognition for Latin characters.[2] Since both dictionaries contain headwords in Rōmaji transcription, one can set up a database relation using the pronunciation transcription as a common key.

2 Japanese character recognition doesn't work well with pre-war orthography, since modern OCR technique concentrates on the most probable characters and prefers contemporary short forms. Further, Japanese OCR has problems with e.g. German umlauts.

Unfortunately, the Japanese language has a lot of homonyms, especially as a result of the intensive borrowing of loan words from Chinese which were adapted to Japanese pronunciation while losing some distinguishing features in the process. For example, the Kimura dictionary gives 38 entries with the pronunciation of *ko* or *kō*.

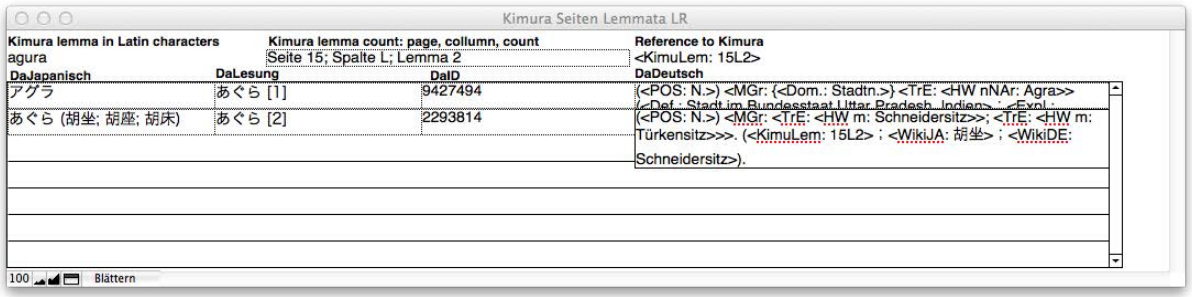


Figure 2: Screenshot of the entry *agura* in a database layout relating to two *WaDokuJT* entries with the same pronunciation.

A relational database being able to deal with homophony – or to be more precise with identical transcriptions of the pronunciation in Latin characters – is shown in Figure 2. The entry with the pronunciation *agura* relates potentially to two entries of the *WaDokuJT* dictionary. The screenshot shows them displayed in a portal of a database layout using a desktop application. A human editor can then enter the reference to the corresponding Kimura entry to the *WaDokuJT* entry via script and a defined hotkey.

A similar approach is used by a students' project at the Institute of Asian Studies – Department of Japanese Studies of Tübingen university in an online version. The project makes serious and steady progress, and, since the first few thousand entries are already covered, it is only a matter of time until the whole dictionary is adapted in this way.

The next step is to give the users access from the *WaDokuJT* webpage to the Kimura information. We will ad a hyperlink to the interface of the online dictionary which will open a scan of the page with the corresponding entry from the Kimura dictionary. This process will be very similar to the one that is explained in Paragraph 4.2, where a hyperlink from the online *WaDokuJT* dictionary opens a page in *Google Books*.

Linking the *WaDokuJT* dictionary with the Kimura dictionary will hopefully also lead to a new correction iteration of the *WaDokuJT* data. As for instance, missing translation equivalents or example sentences from the Kimura dictionary can be added more easily.

3.2 German and Japanese *Wikipedia*

The Japanese *Wikipedia* (2013) is the largest edition of *Wikipedia* in a non-European language. Although the worth of *Wikipedia* as a primary source or reliable references is disputed in academia, it is an easy accessible and important resource. In most cases, it gives a better overview on a certain topic

than would be possible in any ordinary bilingual dictionary. As the *WaDokuJT* dictionary offers mainly translation equivalents and only short hints on meaning or very short definitions, more extensive explanations might be of advantage to many users. *Wikipedia* often refers to corresponding entries in other languages, and Japanese-German pairs can be found, too.

In the source data of the *WaDokuJT* project, the *Wikipedia* article's title is used as unique reference with a mark-up for *Wikipedia* and its language version. As far as possible, *WaDokuJT* refers to both, the Japanese and the German *Wikipedia*. Unfortunately, in many cases a Japanese *Wikipedia* entry has no German equivalent, and often entries, that are marked as corresponding, differ relatively far in meaning and can definitely not be considered as translation equivalents.

The process of finding correspondent *Wikipedia* articles can be automated to a certain extend, although orthographic variety of Japanese makes this process more difficult. Sometimes, several trials with different orthographical forms are necessary. Names of plants and animals for example are written in *katakana* characters within the Japanese *Wikipedia* – as is customary with scientists in the domain of biology – while many traditional dictionaries use *kanji* writings. The *WaDokuJT* project tries to give the most frequent Japanese writing first. Doing so, it is possible to generate automatically links for example to the *Wikipedia* with high probability to work correctly.

Figure 3: Screenshot of the database interface for working on links to the Japanese and German *Wikipedia*.

The screenshot in Figure 3 shows a window from the database application that is used to add links in the *WaDokuJT* dictionary to *Wikipedia* from existing entries. The example entry has four Japanese writings and a generated link to the Japanese *Wikipedia*. A "Web Viewer" field displays the content of this link and enables the editor to check easily the correct correspondence.

A similar link to the German *Wikipedia* can be generated using the first German translation equivalent. If these links between *WaDokuJT* and different *Wikipedia* versions work correctly in spite of possible polysemy or homography, a certain mark-up shorthand as reference to the Japanese or the German *Wikipedia* can be added via script and a hotkey.

The advantage of such a reference for users is, besides getting the translation equivalents which are to be expected from a bilingual dictionary, an easy access to up-to-date encyclopaedic information as well and in many cases even directly both in Japanese and in German.

4 Cooperations with partially open source projects

Currently, the *WaDokuJT* project is cooperating with two domain specific dictionaries, which are the topic of this paragraph. Further cooperation with specialists in other fields is discussed, also. The domains are for example: disaster prevention, mechanical engineering, economy, medicine and life sciences, traditional craft professions and martial arts. Thanks to the great variety and most diverse possibilities, one can be rather optimistic about the results.
The main characteristic of the presented approach is that every project can keep its individuality and its own strengths. The projects shouldn't compete for users or financing, but complement each other in very productive way.

4.1 *Sōgō bukkyō daijiten* – a Buddhist encyclopedia

One cooperating dictionary project is the translation of a Buddhist encyclopaedia (*Sōgō bukkyō daijiten*, 1987; Aoyama et al. 2006; Aoyama et al. 2013). Since a number of years, the articles are translated and extended by information that is relevant for German speaking users. Recently, these data were converted into an online version stored on the same server of Tübingen University as the *WaDokuJT* dictionary.
The entries of the Japanese version of the lexicon are in the Japanese *a-i-u-e-o* order. The online version identifies them by page and count of the entry on the very page. *WaDokuJT* uses this information as unique ID to refer to the Japanese version of the Lexicon as well as to the German translation of the online dictionary. The data of the encyclopaedia will be easier accessible and *WaDokuJT* can concentrate on the translation equivalents and leave buddhological explanations to that lexicon.

4.2 Japanese-German archaeological dictionary and its new Japanese-English-German version

Another collaborating project is an archaeological dictionary project, of which a Japanese-German version already exists as a print-on-demand book: *Kleines Wörterbuch zur Japanischen Archäologie* (Steinhaus, 2010). This project addresses scientific archaeologists and tries to give the latest and often normative translation equivalents, reflecting the current state of the art in the field.

The *WaDokuJT* project has obtained an electronic version of the book from the author and now points from one *WaDokuJT* entry to the corresponding page of the archaeological dictionary. With this information, one can also generate a hyperlink which opens the dictionary page in *Google Books*.

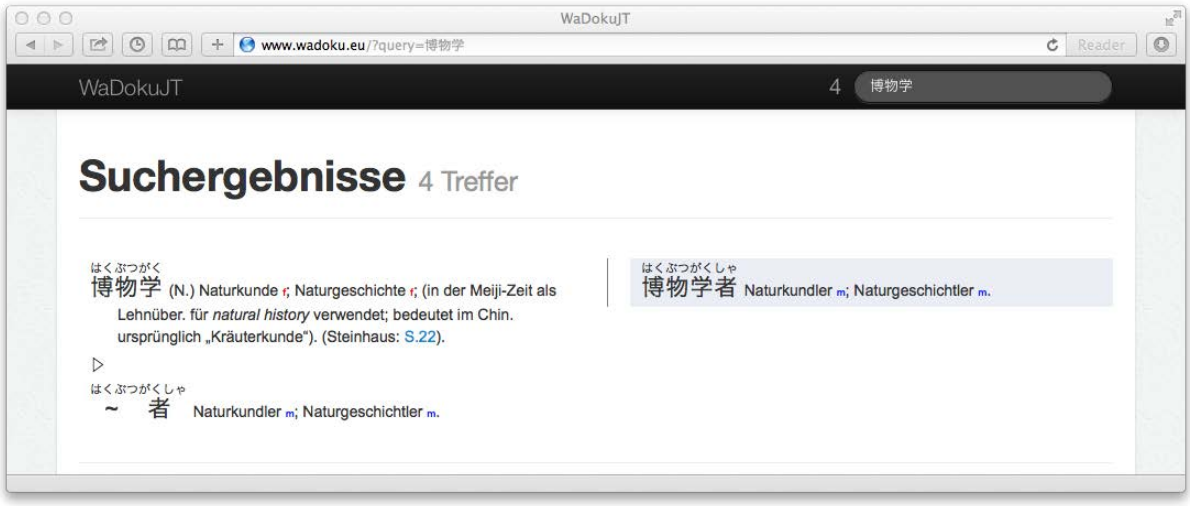


Figure 4: Web interface of WaDokuJT, displaying an entry with hyperlink to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010) at Google Books.

Figure 4 is a screenshot form the online version of the *WaDokuJT* project *wadoku.eu* which shows an entry with reference to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010). The page information is also a hyperlink which by clicking it opens the mentioned page of the book in its online version at *Google Books*.

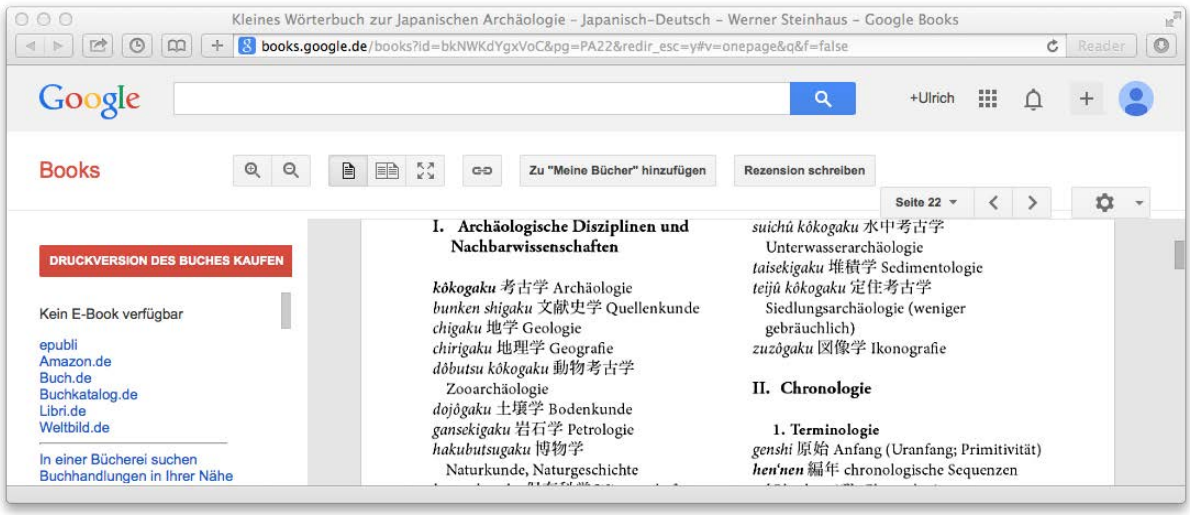


Figure 5: Target page at *Google Books* with the entry *hakubutsugaku* "Naturkunde, Naturgeschichte".

Figure 5 shows a screenshot of the corresponding page of Steinhaus (2010) which is opened at *Google Books*. Here, the translation equivalents are more streamlined for the needs of archaeologists, who

aren't interested in e. g. linguistic explanations about when the word came into usage in Japan and what it meant in its Chinese version.
One main feature of the original archaeological dictionary is the arrangement of entries around certain topics. This means that the archaeological dictionary and the *WaDokuJT* project are no competitors. Further, *WaDokuJT* has no claims of being normative for certain scientific fields but gives possible translations. Through the reference to the archaeological dictionary, users get easy access to the normative translations, too.
The development of the archaeological dictionary goes on in a rather fast pace. In collaboration with the University of East Anglia, Norwich, Great Britain, it is extended to a Japanese-English-German version. In addition, it is planned, that the new version will be hosted at Tübingen as well.

5 Projects linked to *WaDokuJT*

Linking dictionaries to other projects doesn't need to be a one-way-street. Other projects link to the Japanese-German data. For example, *JMdict* (2013), a multilingual lexical database with Japanese as the pivot language uses *WaDokuJT* data for German translations.
Further, a new Korean-German dictionary *Handok.eu* (2013) refers to *WaDokuJT* entries via their ID. Korean and Japanese use a lot of common loan words from Chinese. The project founded by Benjamin Rusch, a former student of the Department of Japanese Studies, may not have the same amount of translation equivalents in the beginning and users may be pleased to get more choice from the *WaDokuJT* project.

6 Managing project derivates

Linking data plays an important role in managing two forks of the *WaDokuJT* project, too. An online interface of the project developed a life of its own, as format changes have rendered it incompatible with the original data. Users can add new entries or suggest corrections of existing entries online. But also on the original data side, new entries and corrections have been added.
Both sides use unique IDs – seven figure IDs for common entries and eight figure numbers for new entries for the online version at *wadoku.de*. Via these IDs entries and changes can be at least monitored, what makes common improvements possible and has most benefit for the users. Data import from *wadoku.de* also includes another edition cycle and should lead to better overall quality of the project.

7 Conclusion

In our opinion, one dictionary alone cannot satisfy all possible users' needs. An attempt to do so may turn the data rather hard to manage and very difficult to use. For additional information, we suggest the approach to link one dictionary to other dictionaries and sources. Our examples come from the daily lexicographical praxis and are added one by one, entry by entry. An automatic system that suggests links, will be helpful for the lexicographer in many cases, but we mistrust a full automatic systems at the moment.

The ultimate aim is to provide users with more information that is better tailored to their specific needs without corrupting usability and manageability of the data.

8 References

Aoyama, T., Paul, G., Schmidt-Glintzer, H., Schmithausen, L. and Wittern, C. (ed.) (2006). *Das Große Lexikon des Buddhismus – Erste Lieferung: A–Bai*. Munich: Iudicium.

Aoyama, T., Paul, G., Rotermund, H. O, Schmithausen, L., Steineck, R. C. and Wittern, C. (ed.) (2013). *Das Große Lexikon des Buddhismus – Zweite Lieferung: Bait–D*. Munich: Iudicium.

Apel, U. (2001): Ein elektronisches japanisch-deutsches Wörterbuch auf Datenbankbasis – Über das Finden von Wörterbucheinträgen im Computer-Zeitalter. In Gössmann, H. and Mrugalla, A. (eds.). *11. Deutschsprachiger Japanologentag in Trier 1999*. Bd. II. Hamburg: Lit. 627–644.

Google Books. http://books.google.com/ [10/11/2013].

Großes Japanisch-Deutsches Wörterbuch (1952). [First edition 1937, Kimura, K. (ed.)]. Tokyo: Hakuyūsha.

Handok.eu. http://handok.eu/ [founded by Benjamin Rusch; 10/11/2013]

JMdict. EDRDG. http://www.csse.monash.edu.au/~jwb/jmdict.html [10/11/2013].

NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen (*NABD*) [draft for DIN 32708, Information und Dokumentation – Umschrift des Japanischen]: http://www.nabd.din.de/projekte/DIN+32708/de/147167717.html [10/11/2013]

Shogakukan Großes Deutsch-Japanisches Wörterbuch (1998): Kunimatsu K. (ed.) [revised edition]. Tokyo: Shōgakukan.

Steinhaus, W. (2010): *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch*. Berlin: Epubli. [Also: on line]. http://books.google.de/books?id=bkNWKdYgxVoC [10/11/2013].

Sōgō bukkyō daijiten (1987): Sōgō Bukkyō Daijiten Henshū Iinkai (ed.). Tokyo: Hōzōkan.

WaDokuJT: http://wadoku.eu/ or http://wadoku.de/ [10/11/2013].

Wikipedia – Die freie Enzyklopädie: http://de.wikipedia.org/wiki/Wikipedia:Hauptseite [10/11/2013]

Wikipedia [Japanese version]: http://ja.wikipedia.org/wiki/メインページ [10/11/2013]

Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch (1989). Schinzinger, R., Yamamoto, A. and Nanbara, M. (ed.): Tokyo: Sanshusha.