Data Documentation and Remote Computing at the International Data Service Center of IZA

N. Askitas[1]

Abstract:
The International Data Service Center (henceforth "IDSC") of the Institute for the Study of Labor (henceforth the "Institute" or "IZA") is the Institute's organizational unit which services its data needs as well as those of its various affiliated communities and the ambient associated research community at large. Since data are at the core of the Institute's makeup, its IDSC (idsc.iza.org) is a major contributor to its vision of a virtual institute. Besides a short presentation of the IDSC we focus on two of its major work areas: data documentation and remote computing.

Keywords: metadata, documentation, remote computing, remore processing, data enclave

1. Introduction

The heritage of IDSC goes back to the 1990's when a group of German economists, including IZA's director K. F. Zimmermann, seeded an intense discussion in Germany about providing the scientific community with a better data infrastructure (R. Hauser, G. G. Wagner and K. F. Zimmermann, 1998). Commissioned by the Federal Ministry of Education and Research, the KVI[2] made several recommendations on how to improve the scientific data infrastructure in Germany (KVI, 2001). These recommendations lead to the establishment of several research and service data centers, among which IdZA (IDSC's predecessor) was the only partner representing labor economics. The German speaking reader may consult (H. Schneider and C. Wolf, 2008) for a good historical write up. After a successful evaluation of the first three-year pilot phase which started in 2003, a second phase of financing was awarded to IZA (2008-2010). The IDSC in its new form represents IZA's continued commitment to the field.

The IDSC is currently undergoing a drastic restructuring of its assets as well as a broadening of its mission. The Center is now building partnerships with other players in the field in order to leverage complementary competence and build mutually beneficial long term relationships and alliances. The main components of the Center's realignment are all contained in the Center's acronym:

1. IDSC is International. It operates without national borders or other artificial frontiers, draws its know-how from the wide international arena and aims at catering to the international research community always in close alignment with the Institute's vision of a virtual institute.

2.IDSC is about Data. This mostly means the technology of data but also involves ethics, legal, educational, and other aspects. IDSC develops, applies, and integrates know-how aimed at dealing with data in the context of operation of IZA. A core component of the work of the IDSC is about inventing, developing, integrating, and deploying/promoting solutions for computing with data and in particular with "difficult", i.e. highly sensitive/confidential, data.

3.IDSC is about Service. IDSC services the data needs of the IZA resident research community, the various global and virtual IZA research communities (Fellow and Affiliate networks, etc.) and the research community at large in that order. The meaning of the order is dual: on the one hand it expresses priority in the sense that IDSC serves the local community first and the remote ones afterwards; on the other hand it expresses deployment order in the sense that local deployment is a preparation step for the large scale deployment. In that sense the IDSC finds itself in the privileged position of belonging to an ideal ambient environment in which to incubate ideas on technology applications.

4. IDSC is a Center. This means that besides being an organizational unit of IZA, it is an entity of its own with relationships to the other IZA units and the world. It is also a center in the sense that it aims to become the focal point of an International, Data-related, Service oriented network of economically-minded technologists and technologically-savvy economists. It also aims to become the ubiquitous place for scientists to look for data support, data access support and data services with emphasis on labor economics but also beyond.

This paper will focus on two important areas of work of the center: data documentation (Section 2) and remote computing (Section 3) and will close with a short mention of future plans and challenges (Section 4).

2. Data Documentation

This work area is deeply rooted into the legacy of the IDSC and represents a significant area of activity in the Center's body of work. Its original purpose was to translate (and along the way standardize) official German metadata in order to improve its usage in the scientific literature. The premise was that contrary to the country's standing as a world-wide top exporting country, the amount of scientific literature based on German official data was rather poor. The IDSC took up the ambitious task of remedying this situation by improving on what it saw as its main cause: the lack of accessible documentation. The metadata offering of the IDSC, which by now goes well beyond translating German metadata into English, consists of a detailed, in depth, searchable and standardized information service, especially helpful for comparative research, which includes an ever growing number of datasets. There are currently datasets in the areas of:

- Employment and Wages
- Education and Training
- Demographics and Migration

For the end user the most important features of the metadata offering of the IDSC in its current implementation (http://idsc.iza.org/metadata/) are:

1. Every dataset has a searchable HTML presentation.
2. Every dataset has its metadata in PDF book form.
3. For every dataset its origin and how to reach it as well as the IZA discussion papers which have used it are included.

For a data professional, it is important to know that the metadata are saved in DDI form (currently version 2) and the DDI files are publicly available for download. Anyone is free to use these DDI files for their own presentation so long as the IDSC is properly cited and the DDI files are made available in the same way. So far as we know the DDI files of the IDSC are currently the only publicly available, variable-level DDI files. The reader may verify how difficult it is to google her/his way to real instances of DDI files anywhere outside IDSC!

Based on these DDI files and using open source and community tools such as the IHSN Microdata Management Toolkit a static HTML presentation is produced for each dataset. Since data documentation is a document like any other and since once compiled it remains largely unchanged the use of relational database supported metadata systems appears to be uncalled for in this context. The most essential element of dynamic implementations is recovered by indexing the static pages so as to make them searchable. This solution is very efficient and performs and scales very well. In what follows the main ingredients of this solution are described by discussing the work process involved in documenting a dataset.

The documentation work typically starts with a DDI file (version 2) which is produced using the Metadata Editor which comes with the IHSN Microdata Management Toolkit. The metadata which flows into such a document is collected in a variety of ways from a variety of sources and data formats all of which are case specific. By using a version of the CD-ROM builder from the ISHN toolkit, HTML and PDF presentations are produced. This version of the CD-ROM builder is modified to include the IDSC branding. Using the keyword attribute of DDI version 2, relevance and context keywords are attached to datasets and to their variables. These keywords belong to a concept hierarchy derived from the HASSET[3]. The keywords are used to make the metadata searchable. The concept hierarchy has a HTML presentation in its own right which is based on its implementation in a relational database. This concept hierarchy module accompanies the searches and maybe used by the user in order to perturb the scope of a search. A search for variables on "Wages" for example will return in addition to the variables that match, the conceptual neighborhood for the concept: synonyms such as "earnings", "pay" or "remuneration" broader concepts such as "income" but also narrower terms such as "low pay" as well as related concepts such "wages policy" all in the form of markup encoded searches accompanied by the number of hits behind the associated search. In effect presented this way the concept hierarchy becomes a kind of directory structure containing relevant results. The search may be restricted to variable pages, variable group pages, dataset overview pages, or dataset dictionary pages. The IHSN toolkit's CD-ROM builder conveniently produces for each dataset: variable pages on which a full description of the variable is presented, group pages on which groups of related variables are presented as well as dictionary or overview pages which summarize the dataset's focus. Making these datasets searchable separately is useful for achieving different ends. Searching dictionary or overview pages maybe used to locate datasets whereas search group and variable pages maybe used to locate variables across datasets.

The indexing and the search of these static pages are done using swish-e (swish-e.org). The indexing occurs once every night: once a dataset is staged by being written into the publishing area of the web service it becomes available to the indexer without further action necessary. A search API is built on top of swish-e which may be used by anyone wishing to integrate a search of the Center's metadata inventory in their own presentations or other data products. An application of the search API is the Stata module of the author (Askitas, 2009) which may be installed in Stata by running "ssc install metadata". The module integrates metadata right into Stata and attempts to bring DDI into the realm of a widely used standard econometric application. An RSS feed enriched with Dublin Core elements is produced programmatically out of the DDI files. This feed is then used to produce a listing of the metadata using Stata's own web capabilities by translating it to SMCL (Stata's own markup language) on the fly. The search API is used to locate datasets relevant to a keyword. Subsequently Dublin Core files (derived from DDI) of these datasets are used to produce the SMCL presentation of the datasets we found. We are planning on integrating sample datasets right into this presentation so the user can search for and program against data in an integrated fashion.

The entire metadata offering of the IDSC is hooked to its own web analytics based on the open source Piwik (piwik.org) so that we can produce access and usage statistics of the metadata offered all the way to the variable level pages.

We plan to integrate metadata into a news aggregator for economics to allow, for example, researchers searching for grants to find the datasets to use in their project proposals. The idea here is to promote metadata right into one of the main activities during which a researcher looks for new data: projects proposals. In order to achieve the merging of news with metadata, news items are tagged with the same keywords as the metadata.

3. Remote Computing

Computing with data has never been as exciting and powerful as it is today. It has also never been as necessary or as ridden with issues and problems. As computing capacity is expanding and large amounts of data can be analyzed faster by empirical data analysts, research projects tend to become increasingly cross sectional and interdisciplinary and this results in more complex computing circumstances. In the past computing capacity and data were collocated in computing centers and there was hence basically one way to compute with it: on site. As computing devices proliferate, their mobility and capacity no longer excluding each other, constant connectivity is becoming the rule and generally computing is democratized this is no longer the case, hence: complex computing circumstances. Researchers can say more and more about the world by means of data based empirical research and they can, want or need to do it with more data from more locations.

On the other hand the demand for empirical research is increasing as the world gains in complexity and is increasingly thought of as a system expressed in equations and measured by variables. In this world privacy, disclosure and data protection acquire a new importance, complexity and perhaps interpretation.

Some of the main stakeholders in this world of data suitable for empirical research are collectors, producers, owners, data custodians and data analysts. Data collectors, producers and owners are usually the side which creates the data. This is where activities such as field work, data reorganization, quality control, definition of access rules etc take place. Data custodians are usually the owners of the data but this is not necessary. Custodianship is transferable by means for contractual agreements. Lastly data analysts are the people who use the data to create knowledge necessary for society to base policy decisions on. In an "ideal world" data is being produced and cared for, custodianship is smoothly and securely regulated and researchers get unlimited access to the data and are able to produce research results without friction. Friction in this setup is created due to the legal requirements, deficits or variability (across countries for example) thereof, the need for privacy protection and disclosure control and the complexity of the computing circumstances. This is the context in which the IDSC's work on computing from afar is taking place.

In Germany the concept of "factual anonymization" is widely accepted since the 90s in effect enabling the creation of so called "scientific use files" which may sometimes be given to the researcher. These files are basically samples of the data which interest researchers and their being "factually anonymized" means that deanonymization is computationally sufficiently expensive. This came as a response to the increasing complexity of the "computing circumstances" mentioned above.

These files however may not be sent abroad and do not always entirely cover a project's data needs. In these cases on site computation is still necessary and this is where the work of IDSC comes into the picture. The Center runs its own data enclave which both conforms to the strictest data security standards and yet strives to achieve the highest possible degree of scientific freedom. To achieve this, IDSC applies a properly stratified way to interface with the scientists working in it:

- Locally through a contained "ultra thin" network segment
- Remotely via several tools in its remote computing portfolio.

Several Research Projects based on highly sensitive datasets are currently hosted within the IDSC Data Enclave. Some of these are:

- "The Long-term Effects of Start-up Subsidies" (M. Caliendo, H.J. Baumgartner, 2008, M. Caliendo and S. Kuenn, 2008, etc)
- "Hartz1b: Evaluation of Further Promotion of Education and Training-Programs" (U. Rinne, A. Uhlendorff and Z. Zhao, 2008; J. Kluve et al, 2007; U. Rinne, M. Schneider, A. Uhlendorff,

2007; H. Schneider, A. Uhlendorff, 2006; H. Bonin, H. Schneider, 2006)

- “Eval5hi: IZA Evaluation Dataset” (G. J. van den Berg, A. Bergemann, M. Caliendo, 2008, etc)
- “Schuleingangsuntersuchung – Einkommensentwicklung und die Gesundheit von Kindern“ based on data from the Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit

The Center offers the possibility for computing from afar whenever it is a lawful data custodian and there are no regulations prohibiting it from doing so. One of the tools used by the center for enabling remote computing is JoSuA which was conceived and developed at the IDSC of IZA. Originally designed in order to grant international researchers access to German labor market data, JoSuA has matured into a flexible data analysis instrument with a configurable degree of automation and is designed to fit the needs and specifications of each individual data provider[4].

JoSuA is therefore suitable for use by data providers who own such data and wish to make it available to a larger research community without jeopardizing the security of their data. Data providers may use JoSuA in one of the following ways:

- Install own JoSuA instance
- Host the data at the IDSC Data Enclave
- Use Hosted JoSuA

In all three flavors JoSuA allows the data provider to maintain full control of the output’s censoring. JoSuA is used by several partners of the IDSC: the IAB (http://fdz.iab.de) which has its own installed instance, currently in test phase; the IQB (http://www.iqb.hu-berlin.de/), which hosts its data at the IDSC Enclave. Discussions with a number of other partners are in progress regarding adoption of JoSuA for their own data and/or research communities. As of this writing JoSuA is undergoing a production streamlining in order to prepare it for a larger install base: versioning, Service Level Agreements, feature management are some of the issues which need to be solved. In the remainder of this section we would like to give a brief summary of JoSuA aimed at both data analysts and data custodians hoping to have more product material to show for soon.

A typical scenario for using JoSuA at the IDSC is as follows. The researcher becomes aware of a dataset in a number of ways: by browsing the Center’s metadata offering, because he/she worked with it during a visit at the institute or through the institute’s fellow network etc. If the IDSC is a rightful host and data custodian of the datataset, it will take up the task of assisting with remote access via JoSuA otherwise (and provided it cannot become a custodian of the data) the Center will mediate between the user and a lawful owner/custodian of the data. The IDSC has a dedicated budget available on a first come first serve basis which covers access costs imposed by the data providers. In case of IDSC custodianship the researcher(s) get JoSuA accounts and are enabled to compute against the dataset(s). The code is submitted via a web interface by either file upload, cut-and-paste, or by email to a specific address. The output is censored to comply with any rules imposed by the data producer and is then released to the user.

JoSuA is agnostic to statistical anonymization or non-disclosure procedures, although it allows for attaching automatic censoring based on either a black list or a white list of incoming commands or outgoing results. Actual control remains with the custodian; JoSuA just facilitates the effective enforcement of existing rules and offers an interface for doing so. JoSuA automates all other aspects of running a service, such as keeping a record of user’s projects and jobs, producing business reports, monitoring performance, and managing the logistics of multiple incoming jobs.

JoSuA is designed to prevent data loss. Malicious user action, malfunction of the product, or other such situations may lead to, at most, contained, non-primary data loss: disclosure of one user’s code or output to other users or to a third party may occur (if for example the user forgets a logged in browser on a foreign workstation) but it will never lead to disclosure of primary data. This is due to the backend architecture of JoSuA which pulls submitted jobs inwards and may not be made to push data outwards. Results may be pushed outwards (manually or programmatically depending on configuration) but these are non-critical in the sense of data disclosure since they are censored and hence publishable.

Generally there are two types of approaches regarding computing with data from afar: remote computing and remote processing. The two types of approaches differ in many ways the most important of which are the degree of interactivity and the degree of exposure of the data. Typically interactivity and data protection are inversely proportional. In Germany the display of data on the screen is regarded as a data transfer (to the location of the screen) and is hence not a viable means of computing with sensitive data from afar if the transfer of the data to the location of the researcher is forbidden. Examples of remote computing tools are products based on the VNC or ICA protocol (owned by RealVNC and Citrix respectively). By remote processing one understands a process in which a remote data analyst submits (e.g. per email) analysis code which is run by a local operator who then returns the possibly censored results to the remote user. JoSuA is neither a remote computing nor a remote processing tool although it has elements of both: it allows more interaction than remote processing and less than a remote computing tool. It support however several concepts such as that of a user community, community operator, project co ownerships etc.

It is important to note that the researcher never gets any direct access to the data. The researcher is allowed to compute against the data and gets to see only publishable results conformal to data owner regulations. This makes JoSuA the only tool applicable in cases commercial packages (which are more interactive and allow the display of data on the screen) are not allowed because the display of the data on the screen is considered "data transfer". JoSuA is suitable even in case where the researcher periodically visits the IDSC or any center running a copy of JoSuA since it allows the researcher to be able to continue working whether locally or from afar.

For researchers who think JoSuA may be helpful for their research and data custodians who would like to look into the possibility of using JoSuA to serve their own data to remote researchers a good way to start is to contact the IDSC at idsc@iza.org or via the IDSC help desk on http://idsc.iza.org/.

5. Future Plans and Challenges

We plan to find an organic, functional way to connect data documentation and remote computing. To that end the work in Askitas, 2009 represents only the beginning and more output is to be expected in this direction. The IDSC is actively working in the areas of data visualization, other forms of data presentation, and new forms of metadata discovery and presentation while actively developing JoSuA further.

More generally the IDSC is involved in a wide array of new and exciting projects and partnerships as its activities are diversifying and its output picking up in both volume and outreach. One of its main challenges will therefore be to integrate a well thought out, commonly accepted framework of legal, ethical, and educational guidelines which will routinely be part of its daily work and operation. This will be done with the assistance of the Institute's interdisciplinary data committee but also of its partners and communities.

References

N. Askitas, The IDSC of IZA, an international data service center, IASSIST08 Technology of Data: Collection, Communication, Access and Preservation, Conference at Stanford University, Stanford, CA, USA May 27-30 2008.

N. Askitas (2009). METADATA: Stata module to enable access to metadata, Statistical Software Components S456988, Boston College Department of Economics.

H. Bonin, H. Schneider (2006). Wirksamkeit der Förderung der beruflichen Weiterbildung vor und nach den Hartz-Reformen, April 2006, published in: Wirtschaftspolitische Blätter 53 (2)

M. Caliendo, H. J. Baumgartner (2008). Turning Unemployment into Self-Employment: Efficiency and Effectiveness of two Start-Up Programs, Oxford Bulletin of Economics and Statistics 70(3).

M. Caliendo, S. Kuenn (2008). Long-Run Effects of Start-Up Subsidies, IZA Discussion Paper No 3880, IZA DP Series.

R. Hauser, G. G. Wagner, K. F. Zimmermann (1998). Memorandum: Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestuetzter wirtschafts- und sozialpolitischer Beratung, Allgemeines Statistisches Archiv 82.

D. Schneider (2008). Fresh Phish (2008), IEEE Spectrum v. 45 no 10.

H. Schneider, A. Uhlendorff (2006). Die Wirkung der Hartz-Reform im Bereich der beruflichen Weiterbildung, Journal for Labor Market Research 39 (3-4)

H. Schneider, C. Wolf (2008). Die Datenservicezentren als Teil der informellen Infrastruktur. In: Rolf, G.; Zwick, M.; Wagner, G.G. (Hrsg.): Fortschritte der informationellen Infrastruktur in Deutschland. (Nomos) Baden-Baden.

J. Kluve, H. Schneider, A. Uhlendorff, Z. Zhao (2007). Evaluating Continuous Training Programs Using the Generalized Propensity Score, IZA Discussion Paper No 3255, IZA DP Series.

KVI, Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001). Wege zu einer besseren informationellen Infrastruktur. (Nomos) Baden-Baden.

U. Rinne, M. Schneider, A. Uhlendorff (2007). Too Bad to Benefit? Effect Heterogeneity of Public Training Programs, IZA Discussion Paper No 3240, IZA DP Series.

U. Rinne, A. Uhlendorff, Z. Zhao (2008). Vouchers and Caseworkers in Public Training Programs: Evidence from the Hartz Reform in Germany, IZA Discussion Paper No 3910, IZA DP Series.

G. J. van den Berg, A. Bergemann, M. Caliendo (2008). The Effect of Active Labor Market Programs on Not-Yet Treated Unemployed Individuals, IZA Discussion Paper No 3825, IZA DP Series.

Notes

[1] N. Askitas, Head IDSC of IZA, Info: http://www.iza.org/

home/askitas, email: askitas@iza.org. This paper is based on the author's talk at the IASSIST08 (N. Askitas 2008).

[2] Commission to improve the informational infrastructure by co-operation of the scientific community and official statistics: http://www.ratswd.de/

[3] The HASSET thesaurus was developed by the UK Data Archive at the University of Essex. Neither the UK Data Archive nor the University of Essex may be held responsible for any errors in this material. We are currently considering other concept hierarchies and taxonomies such as the EUROVOC.

[4] JoSuA is designed conservatively in order to avoid making it vulnerable to recurring internet attacks. For an account of the latest scare with DNS cache poisoning see David Schneider, 2008.