Auswertung des EDaWaX Online Survey on Hosting Options for publication-related Research Data

Hintergrund

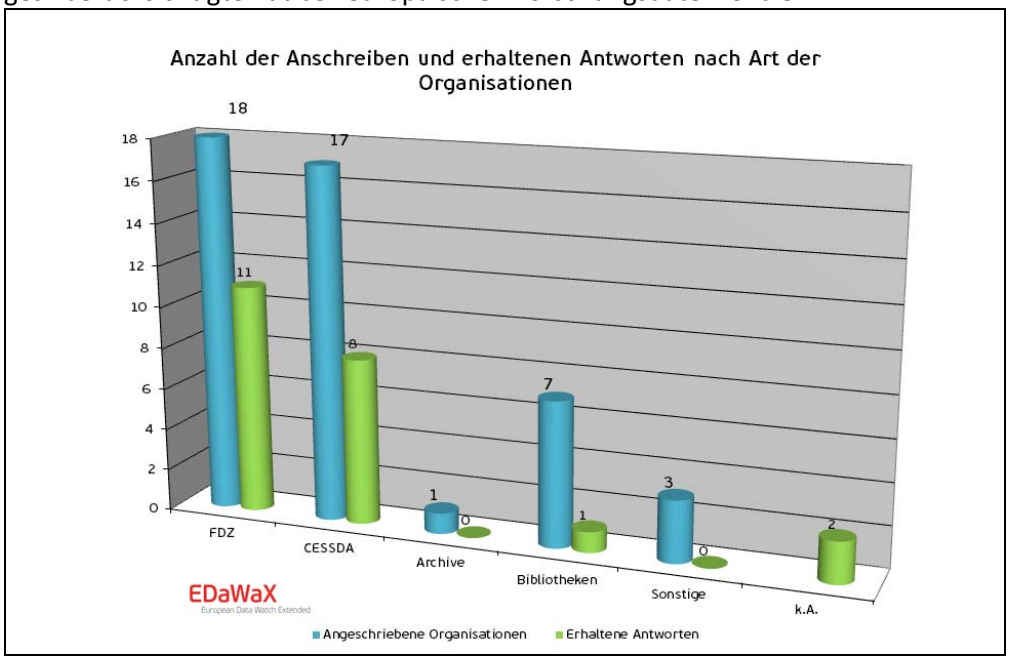
Die vom Projekt EDaWaX („European Data Watch Extended“ [1]) durchgeführte Onlinebefragung verfolgte das Ziel, die vom Rat für Sozial- und Wirtschaftsdaten (RatSWD)[2] akkreditierten Forschungsdaten- und Datenservicezentren, die im CESSDA-Verbund[3] organisierten zumeist europäischen Datenzentren, sowie Bibliotheksverbünde und einzelne Bibliotheken hinsichtlich deren Serviceangebote für die Speicherung und das Hosting von publikationsbezogenen Forschungsdaten zu evaluieren.

Der Aufbau eines solchen publikationsbezogenen Forschungsdatenarchivs für wirtschaftswissenschaftliche Fachzeitschriften ist ein Ziel des DFG-geförderten Projekts EDaWaX.

Die Befragung

Der Onlinefragebogen wurde im Oktober 2012 an insgesamt 46 Organisationen verschickt – darunter 36 nationale und internationale Forschungsdatenzentren (FDZs) und Datenservicezentren (DSZs), 1 Archiv, 7 Bibliotheksverbünde und Bibliotheken sowie drei weitere Institutionen. 22 Organisationen beteiligten sich an der Befragung (47,8%). Die Rücklaufquote ist – gemessen an den Rücklaufquoten schriftlicher Erhebungen – als sehr gut anzusehen.
Bedingt durch die Struktur des Fragebogens beantworteten nicht alle Institutionen jede der gestellten Fragen. Abweichungen der Anzahl der Beantwortungen sind u.a. dadurch zu erklären.

Wichtiger als die Rücklaufquote insgesamt ist freilich die Struktur der Antwortenden bzw. Nicht-Antwortenden. Dabei zeigt sich, dass die allermeisten Antworten (86,4%) aus den deutschen FDZs und SDZs, sowie den Datenzentren des CESSDA-Verbundes kamen. Deutlich unterrepräsentiert sind in den Antworten die deutschen Bibliotheksverbünde und Archive, aber auch die drei unter „Sonstiges“ berücksichtigten außer-europäischen Forschungsdatenzentren.

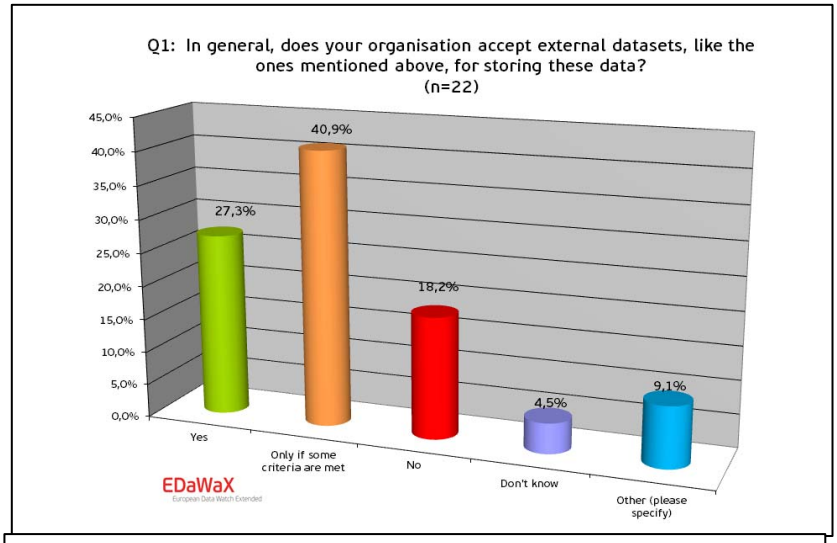


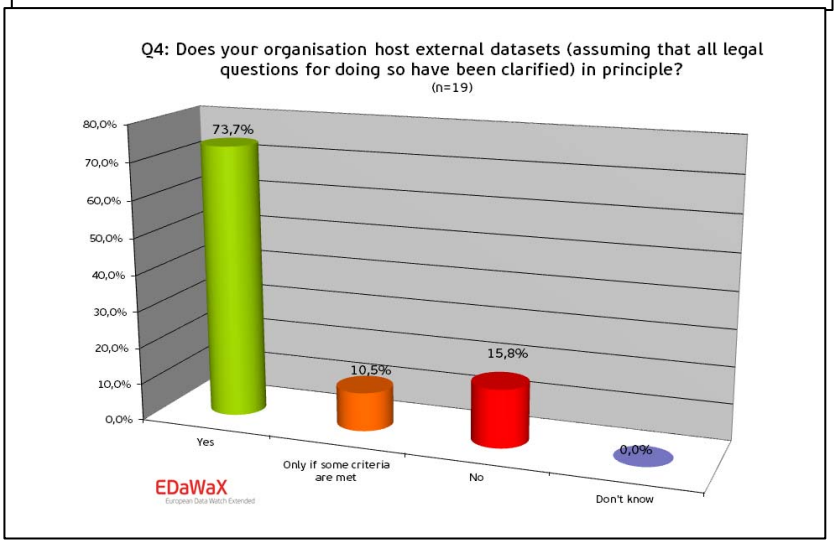
[1] Der Projektblog ist unter der Adresse www.edawax.de erreichbar.
[2] www.ratswd.de
[3] Council of European Social Science Data Archives, www.cessda.org

Für die Bibliotheksverbünde und das angeschriebene Archiv kann nur gemutmaßt werden, dass keine entsprechenden Services oder Angebote zum Bereich Forschungsdaten bestehen und daher eine Beantwortung des Fragebogens aus diesem Grunde nicht erfolgte.

Inhaltliche Auswertung

Zunächst wurde mit der Befragung untersucht, ob die angeschriebenen Institutionen publikationsbezogene Forschungsdaten speichern und hosten. Zudem wurde ermittelt, ob auch selbstgeschriebene Software und der Berechnungscode von statistischen Auswertungen durch die Institutionen gespeichert und bereitgestellt werden. All diese Daten sind oftmals Bestandteil empirischer Forschung in den Wirtschaftswissenschaften.

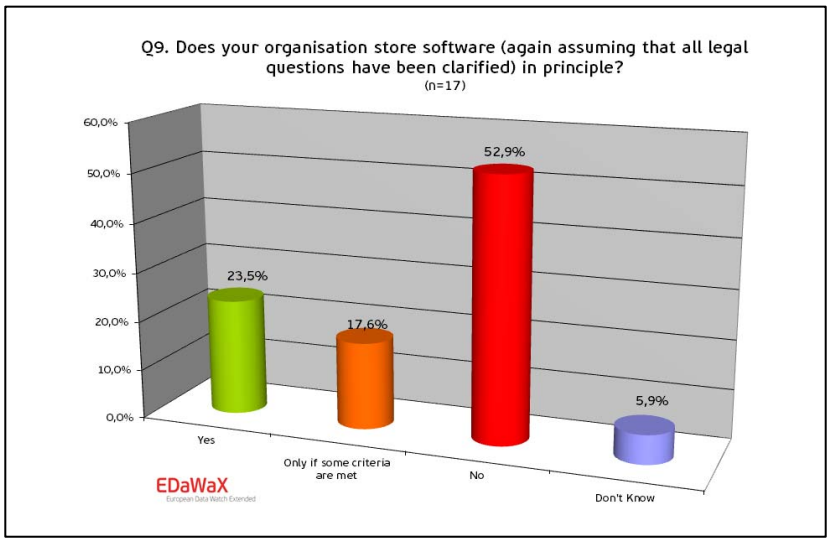




Datensätze

Von den untersuchten Organisationen akzeptierten mehr als drei Viertel externe Datensätze zur Speicherung. Der größte Anteil der Befragten gab dabei an, solche Forschungsdaten nur anzunehmen, wenn gewisse Kriterien erfüllt sind. Solche Kriterien bestehen etwa in Form der spezifischen thematischen Zuständigkeiten der FDZs, aber auch in Form von regionalen, überregionalen und/oder fachspezifische Zuständigkeiten. Zudem wurden technisch-organisatorische Aspekte (Dokumentation, Maschinenlesbarkeit) und rechtliche Fragestellungen als Kriterien genannt.

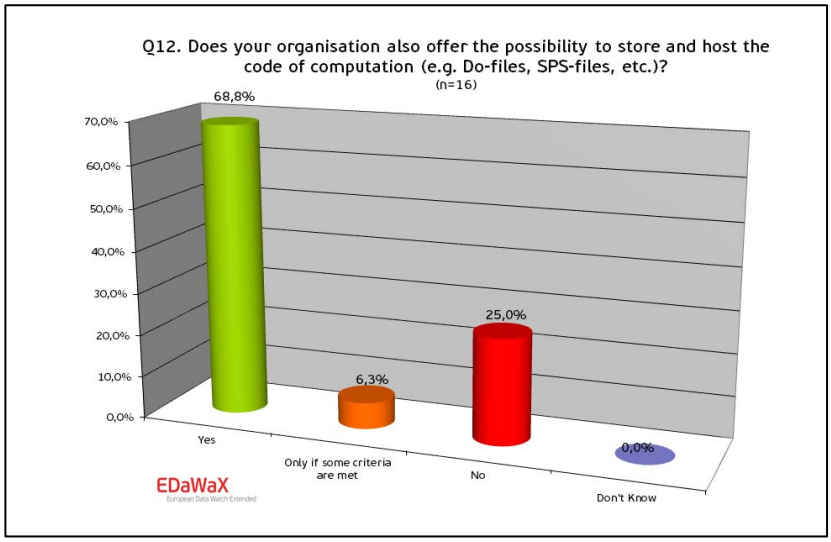
Von den befragten Organisationen gaben etwa 74% an, solche Forschungsdaten auch zu hosten. Wenn dafür Kriterien bestehen, wurde erneut vor allem die fachliche Ausrichtung der Institution als Kriterium genannt.



Software

In Bezug auf Speicherung und Hosting von (selbstgeschriebener) Software, wie sie etwa im Rahmen wirtschaftswissenschaftlicher Simulationen Ver-

wendung findet, zeigte sich, dass nur eine Minderheit von knapp einem Viertel der untersuchten Organisationen eine Speicherung und das Hosting solcher Daten ohne Einschränkungen unterstützt. Weitere 17% betonten, dass auch für solche Software Kriterien existieren, wonach entschieden wird, ob Speicherung und Hosting erfolgen (beispielsweise, wenn diese wichtig für die Datenanalyse ist). Einige Organisationen gaben an, dass entsprechende Services für die Zukunft geplant seien oder solche Software als Teil der Dokumentation zu Datensätzen bereitgestellt wird.
Somit sind die Speicherung und das Hosting solcher Software als „Gaps" anzusehen, die bislang nur eine übersichtliche Zahl an Organisationen anbietet.

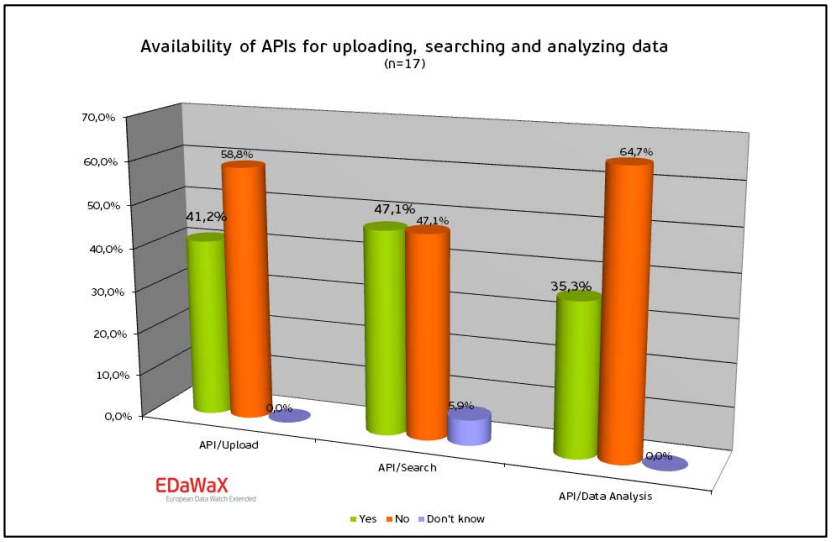


Syntax

Fast 70% der untersuchten Organisationen bieten die Möglichkeit Syntax zu speichern und bereit zu stellen – ¼ der untersuchten Organisationen tut dies jedoch nicht und plant dies auch nicht für die Zukunft. Ein Befragter nannte zudem das Kriterium, dass dies nur bei *derived variables* sinnvoll sei.

APIs

Im Zuge unserer Untersuchung wurde auch die Verfügbarkeit von Programmierschnittstellen (so genannten APIs) thematisiert, mit denen ein automatisierter Austausch von Daten ermöglicht wird.

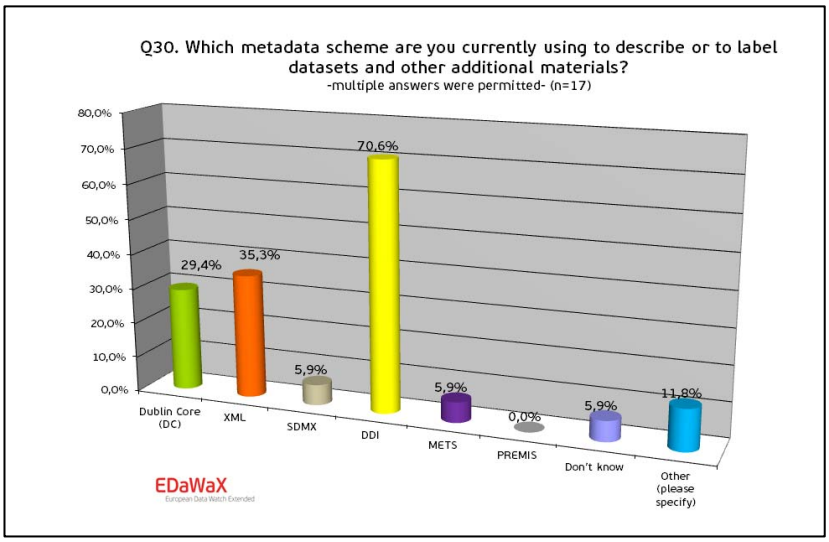


Unsere Befragung ergab, dass weniger als die Hälfte aller Organisationen angaben, über solche Schnittstellen zu verfügen. Am häufigsten wurden APIs zur Suche von Datensätzen genannt (47%), gefolgt von APIs für den Upload von Forschungsdaten. Etwas mehr als ein Drittel (35%) der Befragten gibt zudem an, über Schnittstellen für die Analyse von Forschungsdaten zu verfügen.
Eine durch das EDaWaX-Projekt erfolgte Prüfung der angegebenen Schnittstellen ergab jedoch, dass es sich bei diesen Schnittstellen durchgängig nur um Suchmasken auf den Webseiten handelt. Schnittstellen im Sinne eines extern möglichen Lese- und Schreibzugriffs sind daher als weitgehend unbekannt zu charakterisieren.

Metadaten und Metadatenerstellung

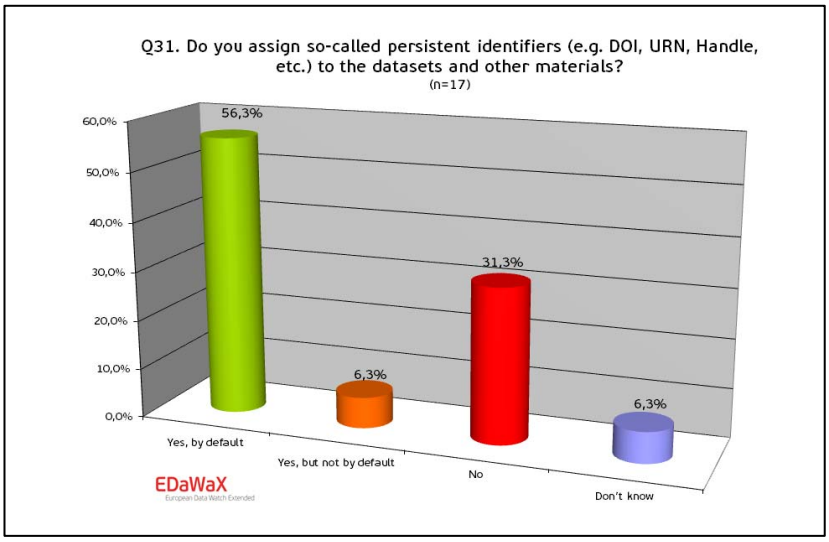
Genutzte Metadatenschemata

Wir interessierten uns zudem für die Metadatenschemata die die Organisationen bei ihrer Arbeit verwenden. Dabei zeigte sich, dass mehr als 70% der Befragten DDI verwenden. Wesentlich seltener wurde XML (35%) oder DC (knapp 29%) genannt. Alle anderen Metadatenschemata werden nur vereinzelt genutzt.



Persistente Identifier (PI)

Fraglich war zudem, ob persistente Identifikatoren (wie handle, DOI, URN, etc…) in den Organisationen Verwendung finden. Die persistente Identifikation von Forschungsdaten ist u.a. wichtig für die Zitierbarkeit von Forschungsdatensätzen. Organisationen aus unserem Sample vergaben solche Identifier standardmäßig in mehr als 56% der Fälle, fast ein Drittel vergab solche Identifikatoren jedoch nicht.

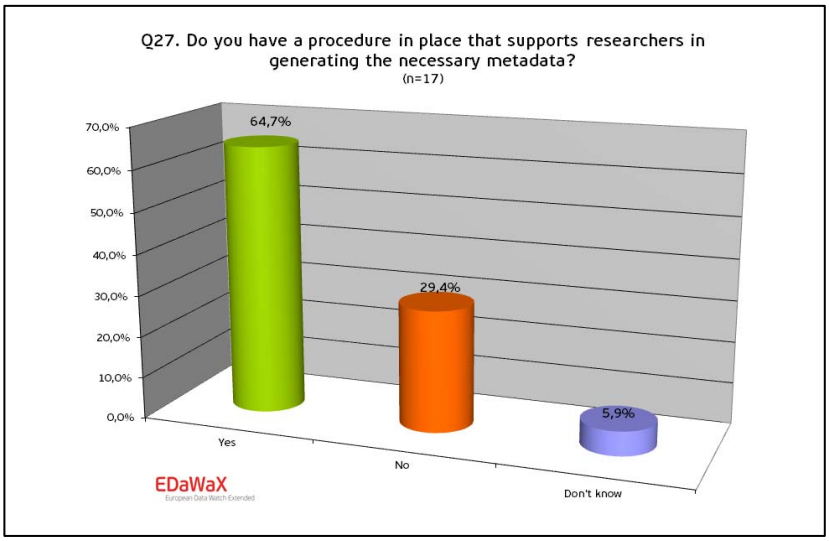


Unterstützung von Semantic Web Technologien

In unserer Befragung wurde auch nach der Verwendung von RDF gefragt – einem Datenmodell zur Beschreibung von Ressourcen mittels semantischer Technologien. Von den befragten Organisationen gab nur eine Minderheit von etwa 6% an, RDF-Daten bereit zu stellen. Fast ¼ der Befragten machte hierzu keine Angaben, was teilweise auf einen geringen Bekanntheitsgrad von RDF zurückzuführen sein dürfte.

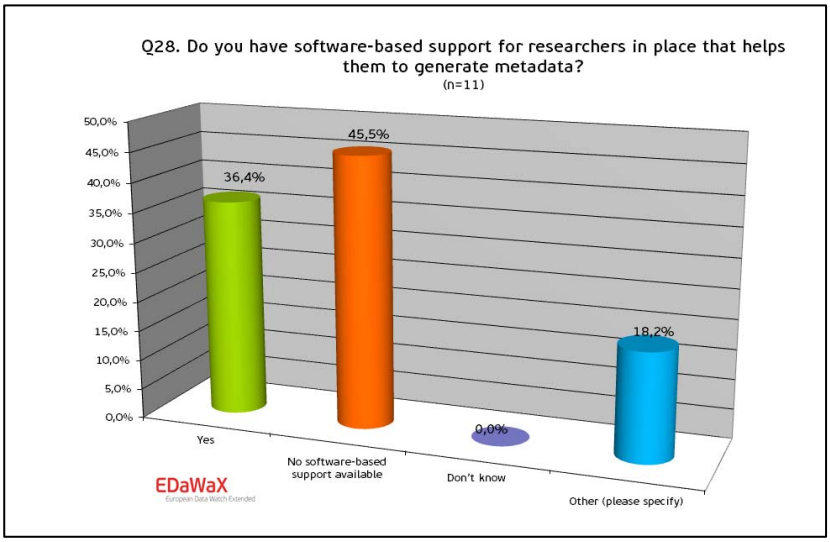
Support bei der Metadatenerstellung

Die Achillesferse für die Nachnutzbarkeit von Forschungsdaten ist häufig die Qualität der Dokumentation von Forschungsdaten. Daher war es von Interesse zu erfahren, ob und wie die untersuchten Organisationen Forscher/innen bei der Generierung von Metadaten unterstützen.



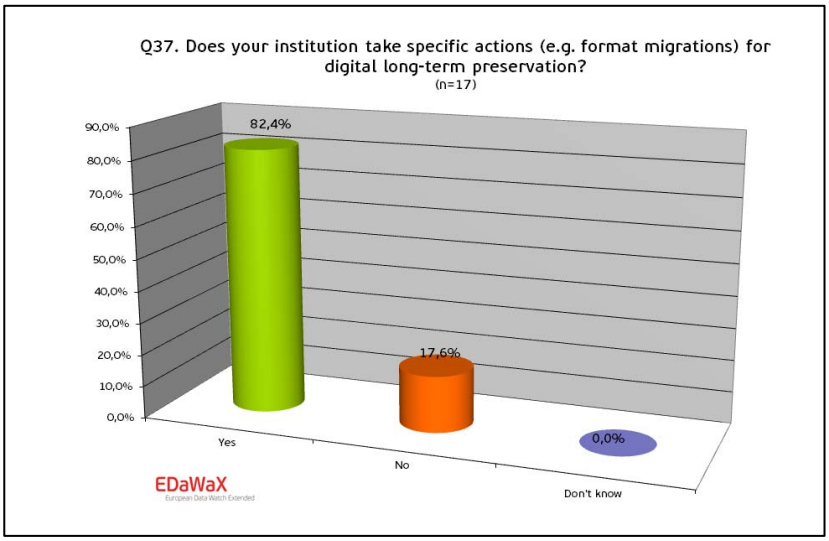
Unsere Befragung ergab, dass

die Mehrheit der Organisationen (fast 65%) über einen entsprechenden Support für Wissenschaftler/innen verfügt.



Dabei interessierte uns auch, ob es softwarebasierte Unterstützung für die Erstellung von Metadaten an den Institutionen gibt, wie dies beispielsweise durch entsprechende Eingabemasken und die Konvertierung der eingegebenen Inhalte in standardisierte Metadaten der Fall sein kann.
Hier zeigte sich, dass über 35% der Befragten über eine solchen softwarebasierten Support verfügen. Auffällig ist die Zahl an Nennungen im Bereich *other*. Hier wurden beispielsweise auch schriftlich auszufüllende *Data Deposit Forms* aufgeführt.
Unsere Nachfrage nach dem Namen bzw. der Art der Software ergab, dass mindestens zwei Institutionen Nesstar[4] einsetzen. Viele Organisationen nutzen zudem Eigenentwicklungen.

Langzeitarchivierung
In unserer Befragung wollten wir auch wissen, in welchem Umfang Maßnahmen zur Langzeitarchivierung von Forschungsdaten getroffen werden. Hier zeigte sich, dass mehr als 80% der Befragten entsprechende Maßnahmen umsetzen.



Fazit:

Die Befragungsergebnisse zeigen, dass Datenzentren ein relevanter Speicherort für publikationsbezogene Forschungsdaten sein können, da sie verschiedene Voraussetzungen dafür bereits erfüllen. Dennoch gibt es unter den befragten Organisationen bislang keine Institution die in Gänze alle Anforderungen hinsichtlich Speicherung und Hosting solcher publikationsbezogenen Forschungsdaten erfüllt.

Im Einzelnen ergaben sich folgende Ergebnisse:

- Etwa Dreiviertel aller befragten Einrichtungen akzeptieren grundsätzlich externe Forschungsdaten, inklusive publikationsbezogener Forschungsdaten. Allerdings gibt es z.T. Einschränkungen, etwa aufgrund der fachlichen oder regionalen Zuständigkeit oder hinsichtlich der qualitativen Anforderungen an solche Datensätze.
- Fast ebenso hoch (annährend 75%) ist die Anzahl der Datenzentren, die den zugehörigen Berechnungscode (Syntax) der abgegebenen Berechnungen prinzipiell speichern und hosten. Falls zur Berechnung empirischer Ergebnisse spezielle (selbstgeschriebene) Software ver-

[4] www.nesstar.com

wendet wurde, wird diese allerdings nur von etwa 40% der Befragten für Speicherung und Hosting akzeptiert.

- An eingesetzten Metadatenschemata dominiert klar DDI (70%) vor XML und Dublin Core (35 bzw. 30% - Mehrfachnennungen möglich). Knapp zwei Drittel zeichnet die Datensätze zudem mit Persistenten Identifikatoren aus und macht sie so leichter zitierfähig. Etwa Dreiviertel aller Befragten leistet zudem Unterstützung bei der Eingabe der Metadaten durch Forschende.
- Schnittstellen für die externe Suche oder den Upload von Datensätzen werden bislang nicht durch die befragten Einrichtungen angeboten. Kaum verbreitet ist auch der Einsatz von semantischen Technologien wie z.B. RDF.
- Die befragten Organisationen sorgen zudem zu über 80% für die Langzeitverfügbarkeit der bei ihnen gehosteten Datensätze.