WissGrid

Deliverable 2.3

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Arbeitspaket 2: Blaupausen und Beratung,
Arbeitspaket 3: Langzeitarchivierung

Checkliste zum Forschungsdaten-Management[1]

Version 0.6:
Entwurfsversion zur öffentlichen Kommentierung

[1] This work is created by the WissGrid project. The project is funded by the German Federal Ministry of Education and Research (BMBF).

Herausgegeben von

WissGrid – Grid für die Wissenschaft
Teil des D-Grid Verbundes und der deutschen e-Science Initiative
www.wissgrid.de

BMBF Förderkennzeichen: 01|G09005A-G (Verbundprojekt)
Laufzeit: Mai 2009 - April 2012
Dokumentstatus: Deliverable

Kontakt

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Jens Ludwig
Abteilung Forschung & Entwicklung
Papendiek 14
37073 Göttingen

Leibniz-Institut für Astrophysik Potsdam
Harry Enke
Abteilung e-Science
An der Sternwarte 16
14482 Potsdam

Autoren

Harry Enke, Leibniz-Institut für Astrophysik Potsdam
Norman Fiedler, Institut für Deutsche Sprache Mannheim
Thomas Fischer, Niedersächsische Staats- und Universitätsbibliothek Göttingen
Timo Gnadt, Niedersächsische Staats- und Universitätsbibliothek Göttingen
Erik Ketzan, Institut für Deutsche Sprache Mannheim
Jens Ludwig, Niedersächsische Staats- und Universitätsbibliothek Göttingen
Torsten Rathmann, Deutsche Klimarechenzentrum
Gabriel Stöckle, Zentrum für Astronomie der Universität Heidelberg

WissGrid, 2011

Vorwort - Anleitung zur Benutzung der Checkliste

Um digitale Forschungsdaten langfristig nutzen zu können und den Anforderungen an die gute wissenschaftliche Praxis gerecht zu werden, ist es notwendig, eine Reihe von Vorkehrungen und Maßnahmen zu treffen. Ohne ausreichende Dokumentation und Metadaten können Forschungsdaten nicht verstanden und verwaltet werden, ohne Finanzierung und Personal können sie nicht aufbewahrt und gepflegt werden und ohne definierte Arbeitsabläufe für das Datenmanagement können nur geringe Datenmengen mit unklarer Qualität gesichert werden. Das sind nur wenige Beispiele der Aufgaben, die idealerweise schon bei der Planung eines Forschungsvorhabens berücksichtigt werden sollten.

Diese Checkliste soll Vorhaben helfen, zusammen mit Infrastruktureinrichtungen wie Rechenzentren oder Datenarchiven den Umgang mit Forschungsdaten zu planen, wie es z.B. in DFG-Anträgen gefordert ist. Einführungen in die einzelnen Themengebiete und weiterführende Literaturhinweise finden sich in dem Leitfaden des WissGrid-Projekts zur Langzeitarchivierung von Forschungsdaten. Einen Überblick der Aufgaben des Forschungsdatenmanagements geben die Abbildungen 1 (S. 4) sowie 2 (S. 9).

Formale Informationen zur Checkliste

a) Erstellungsdatum

b) An Erstellung beteiligte Personen

Teil I

Aufgaben im Lebenszyklus von Forschungsdaten

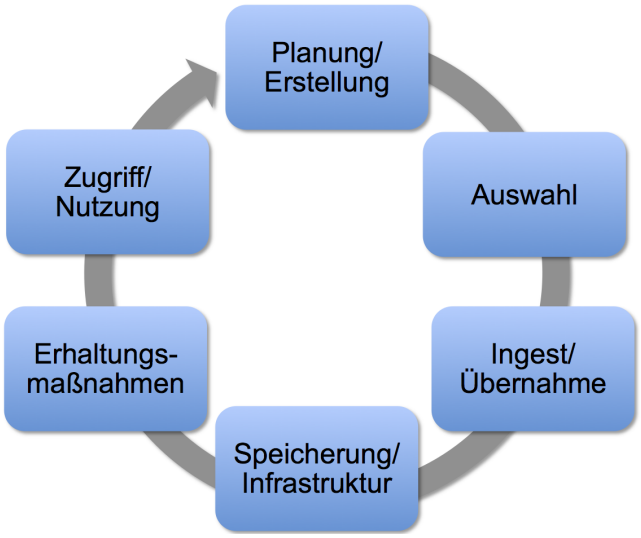

Abb. 1: Der Lebenszyklus von Forschungsdaten

1 Planung und Erstellung

1.1 Allgemeine Angaben zu den Rahmendaten des Vorhabens

a) Wie lautet der Name/die Bezeichnung des Forschungsvorhabens?

b) Was sind die Ziele des Projekts?

c) Wer sind der/die Projektträger oder Finanzgeber?

d) Was ist die angestrebte/genehmigte Laufzeit des Projekts?

e) Welche Organisationen sind an dem Forschungsvorhaben beteiligt (Projektpartner)?

f) Welche Organisation oder Person fungiert als Projektverantwortlicher/Leiter/Koordinator?

1.2 Vorhandene Daten

a) Können bereits existierende Datensätze benutzt werden? Wurde nach Datenbeständen im Besitz der eigenen Institution und von Dritten recherchiert?

b) Welche Bedeutung haben die vorhandenen sowie die erzeugten Daten für das Vorhabensziel? Wieso sind die Daten wichtig? *(z.B. Dokumentation, Publikation, Nachnutzung, . . .)*

c) Wie wird die Integration zwischen den bereits bestehenden und den neuen Daten organisiert? Wie wird z.B. Herkunft und Qualität der Daten dokumentiert?

1.3 Arten von Daten

a) Welche Datenarten werden verwendet bzw. erzeugt? *(z.B. Beobachtungsdaten, Simulationsdaten, Video-Interviews, . . .)*

b) Inwieweit sind die Daten reproduzierbar?

c) Wie werden Daten erfasst oder erstellt? *(z.B.: welche Instrumente, Technologien und Verfahren werden benutzt und anhand welcher Kriterien wird entschieden, ob ein Datensatz erzeugt wird?)*

d) Wie groß ist die geschätzte Datenmenge/Produktionsrate?

e) Welche Maßnahmen werden zur Qualitätssicherung bzw. für das Qualitätsmanagement ergriffen? *(z.B. Dokumentation, Kalibrierung, Validierung, Überwachung, Transkriptionsmetadaten, peer-review)*

f) In welchen Dateiformaten werden die Daten vorliegen? Mit welchen Datenformaten wird gearbeitet?

2 Auswahl und Aufbewahrungsdauer

2.1 Gründe zur Aufbewahrung

Wieso müssen welche Daten ganz oder teilweise aufbewahrt werden? *(Mehrere Antworten möglich)*

a) Arbeitskopie: Sollen die Daten für die aktive Arbeit während des Vorhabens gesichert werden?

b) Nachweis der guten wissenschaftlichen Praxis: Sind die Daten Grundlage einer Publikation?

c) Nachnutzung: Sind die Daten für spätere Forschung wichtig und nicht effizient reproduzierbar?

d) Auflagen: Unterliegen die Daten rechtlichen oder vertraglichen Auflagen bzgl. ihrer Aufbewahrung? Welchen?

e) Dokumentation: Sind die Daten gesellschaftlich bzw. politisch relevant?

2.2 Datenauswahl

a) Zu welchem Zeitpunkt erfolgt die Selektion?

b) Wer ist für die Auswahl verantwortlich?

c) Welche Hilfsmittel (z.B. Software) werden für die Selektion verwendet?

d) Welche Kriterien werden für die Auswahl festgelegt?

2.3 Aufbewahrung

a) Wie lange sollen welche Daten aufbewahrt werden? *(z.B. bis zum Ende des Projekts, 10 Jahre nach Ende des Projekts, bis zu einem bestimmten Ereignis, unbefristet, . . .)*

b) Wie ist das Verfahren, wenn Daten ggf. nicht mehr aufbewahrt werden sollen? Werden die Produzenten benachrichtigt?

3 Ingest: Einspeisen und Verantwortungsübernahme

3.1 Verfahren

a) Wann werden die Daten übergeben (in welcher Projektphase und wann innerhalb des Ingest-Workflows)?

b) Wie werden die Daten übertragen?

c) Wann und von wem werden welche Metadaten erfasst?

d) Wie werden Daten und Metadaten auf technische/formale Korrektheit und Vollständigkeit überprüft (Validierung)?

e) Wie werden sensible Daten behandelt?

3.2 Verantwortungsübernahme

a) Sind die Rechte und Pflichten der Datenproduzenten und des Datenarchivs/Repository geklärt? Wer ist für welchen Teil des Ingest verantwortlich?

b) Gibt es eine Übernahmevereinbarung?

c) Wird der Ingest protokolliert?

d) Ist ein Vorgehen für die Fehlerbehandlung definiert?

4 Speicherung und Infrastruktur

4.1 Datensicherung

a) Wer ist während des Projekts und wer ist nach dem Projekt verantwortlich für die Speicherung der Daten?

b) Mit welchen Technologien werden die Daten gespeichert?

c) An welchen Orten werden die Daten gespeichert?

d) Werden regelmäßig zusätzliche Sicherheitskopien erstellt und überprüft?

4.2 Infrastruktur

a) Wie hoch wird die erwartete Datenmenge sein (pro Jahr oder in der Gesamtdauer des Projekts)?

b) Welche Netzwerk-Bandbreite ist für den Datentransfer und Zugriff erforderlich?

c) Welche Formen des Zugriffs sind vorhersehbar? Wie häufig und intensiv wird auf die Daten zugegriffen? Müssen die Daten online sein, oder reichen auch nearline- oder offline-Speicher (Band)?

d) Gibt es spezielle Anforderungen durch besondere Dienste zur Datennutzung? *(z.B. Grafikkarten bei Visualisierung, Rechenkapazität für Datenextraktion, . . .)*

5 Erhaltungsmaßnahmen und ihre Planung

a) Sind die eingesetzten Technologien und Abhängigkeiten von anderen Datensätzen oder Diensten dokumentiert?

b) Sind die Nutzungszielgruppe und die Anforderungen an die Nutzung der Daten dokumentiert?

c) Wird regelmäßig überprüft, ob sich diese Anforderungen sowie die verfügbaren Technologien oder Abhängigkeiten verändert haben?

d) Gibt es eine Neubewertung der Aufbewahrungswürdigkeit nach einem definierten Zeitraum/Ereignis?

e) Gibt es eine Nachfolgeregelung für den Fall, dass die aufbewahrende Institution die Aufgabe abgeben muss?

6 Zugriff und Nutzung

6.1 Nachnutzung und Suchbarkeit

a) Können prinzipiell die Daten auch von Anderen innerhalb oder außerhalb des Projekts genutzt werden?

b) Gibt es Gründe, die Daten prinzipiell nicht zu freizugeben? *(z.B. Datenschutz, Geheimhaltung etc.)*

c) Gibt es eine Verpflichtung, die Daten freizugeben? *(z.B. durch den Geldgeber)*

d) Welche Einrichtungen bzw. Gruppen werden voraussichtlich an den Daten interessiert sein?

e) Wie ist das Verfahren, um Zugriff auf die Daten zu bekommen?

f) Sind die Daten suchbar?

g) Wie werden die Daten veröffentlicht bzw. bekannt gemacht?

h) Was sind die beabsichtigten oder vorhersehbaren Verwendungen der Daten? Mit welchen Diensten/Programmen werden die Daten üblicherweise genutzt?

6.2 Offener Zugang versus Zugriffsbeschränkungen

a) Gibt es ein Recht auf Erstnutzung durch den Ersteller der Daten? Ab wann dürfen auch Andere auf Daten bzw. Metadaten zugreifen (Sperrfristen)?

b) Unterliegen die Daten Nutzungseinschränkungen oder Lizenzbedingungen?

c) Wird der Zugang gebührenpflichtig sein?

d) Wie werden die Zugriffsbedingungen durchgesetzt und technisch implementiert? Wer ist für die Durchsetzung verantwortlich?

6.3 Interoperabilität

a) Sind fremde Dienste oder Archive interoperabel zu eigenen Datendiensten?

b) Auf welchen Ebenen/in welcher Hinsicht wird Interoperabilität gefordert/gewährleistet/angestrebt?

Teil II

Übergreifende Aufgaben des Forschungsdaten-Managements

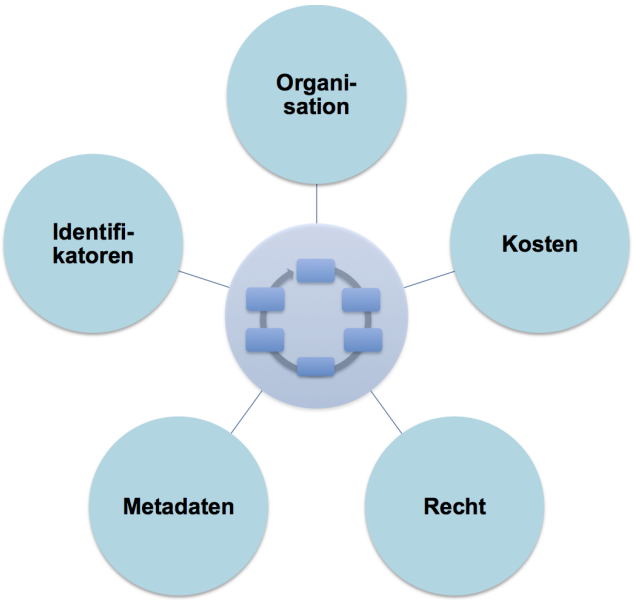

Abb. 2: Übergreifende Aufgaben des Forschungsdaten-Managements

7 Organisation, Management und Policies

7.1 Organisation und Management

a) Werden die Daten in einem institutionellen, kollaborativen oder thematischen Repository aufbewahrt?

b) Ist der Bezugsrahmen national oder international? Erwachsen daraus besondere Anforderungen?

c) Welche Organisationseinheit ist zuständig für das Datenmanagement?

d) Welche Institutionen sind am Datenmanagement beteiligt?

e) Sind die beteiligten Institutionen, Organisationen, Personen benannt und informiert? Sind deren Beiträge definiert? Liegt deren Einwilligung vor?

f) Ist der Workflow für das Datenmanagement beschrieben?

g) Liegt eine Beschreibung und Abschätzung von Personal- und anderen Ressourcen vor?

7.2 Policies

a) Welche Anforderungen und Vorgaben bestehen
   - von Seiten der Finanzgeber,
   - von den beteiligten Institutionen oder Forschungsgruppen,
   - von der Fachgemeinschaft,
   - von relevanten internationalen Vorhaben
   - von anderen Seiten

   zum Datenmanagement der Daten?

b) Welche Auflagen liegen ggf. von Seiten eines Datengebers liegen vor?

c) Welche Policies müssen für die Datenaufnahme, den Betrieb und die Nutzung des Repository erstellt/durchgesetzt werden? Wer ist hierfür verantwortlich?

d) Welche Selektionsverfahren für die Daten sind definiert? Gibt es Prozesse innerhalb der Fachgemeinschaft, die zur Selektion der Daten dienen können?

7.3 Einhaltung der Vorgaben und Planung

a) Wie, wann und von wem wird die Einhaltung der Planung überprüft oder nachgewiesen?

b) Wie, wann und von wem wird diese Checkliste bei Bedarf aktualisiert?

8 Kosten

8.1 Kosten- und Aufwandsabschätzung

a) Sind die Kosten und der Personalaufwand abgeschätzt worden? *(z.B. anhand von Vergleichswerten, einer Lebenszyklus-Analyse, etc.)*

b) Welche Kosten entstehen während der Projektlaufzeit und welche danach?

c) Wer übernimmt die Kosten? Wie hoch ist das im Projekt veranschlagte Budget für Datenmanagement?

8.2 Anreize

a) Sind den Beteiligten die Gründe für das Management von Forschungsdaten klar? *(Siehe Frage 2.1)*

b) Ist es notwendig, den individuellen und allgemeinen Nutzen herauszuarbeiten oder Anreizsysteme einzuführen?

c) Sind die Kosten für das Datenmanagement ins Verhältnis zu den Kosten der Datenerzeugung gesetzt worden?

9 Rechtliche Aspekte von Forschungsdaten

Die rechtliche Absicherung erfordert unter Umständen das Hinzuziehen von Fachleuten wie z.B. Juristen. Es ist jedoch notwendig, eine angemessene Balance zwischen rechtlicher Absicherung und pragmatischer Forschungspraxis zu finden. Untätigkeit in diesen Fragen aufgrund gesetzlicher Überforderung soll vermieden werden.

9.1 Datenschutz - personenbezogene Daten

a) Unterliegen die Forschungsdaten dem Datenschutz? Handelt es sich bei den Forschungsdaten um „personenbezogene Daten“ im Sinne des Bundesdatenschutzgesetz (BDSG) §3 ?

b) Welche Anstrengungen wurden unternommen, um den Anforderungen des Datenschutzes zu genügen?

c) Gibt es ethisch, kommerziell oder in anderer Hinsicht sensible Daten?

d) Welche Maßnahmen werden zum Schutz dieser Daten getroffen?

9.2 Urheberrecht

Die Frage des Urheberrechts an Forschungsdaten ist juristisch nicht abschließend geklärt, und auch das Urheberrecht selbst unterliegt Veränderungen. Es ist sinnvoll, diese Fragen mit den Rechteinhabern explizit zu klären bzw. Urheber- und Nutzungsrechte vertraglich zu regeln.

a) Werden fremde Forschungsdaten oder Software verwendet, welche dem Urheberrecht, dem Patentrecht oder anderen geistigen Eigentumsrechten unterliegen? Wenn ja, wer besitzt die Rechte?

b) Unterliegen eigene Forschungsdaten oder Software dem Urheberrecht, dem Patentrecht oder anderen geistigen Eigentumsrechten? Wenn ja, wie werden sie lizenziert und welche Nutzungsrechte werden eingeräumt? *(z.B. Einschränkungen oder Verzögerungen der Datenverfügbarkeit)*

c) Wenn entsprechende Rechte an den Forschungsdaten bestehen, werden alle notwendigen Maßnahmen zum Zwecke des Datenmanagements eingeräumt?

d) Existieren Schutzfristen für die Forschungsdaten, welche während des Aufbewahrungszeitraumes enden? Wie wird mit diesen Daten verfahren?

10 Metadaten

a) Welchem *Zweck* dient die Einrichtung eines Metadatensystems: wozu sollen die Metadaten dienen oder benutzt werden? *(z.B.: Daten sichtbar machen, Interpretation, Austausch, Verwaltung/Pflege, Präsentation)*

b) Welche *Informationen* sollen durch Metadaten beschrieben werden? *(z.B. Objekte, Akteure, Quellen, Vorgänge, Ergebnisse)*

c) In welchem Rahmensystem werden die Metadaten erfasst, welche Bedeutungen sollen bzw. können darin abgebildet werden? Welcher *Semantik* unterliegen die Metadaten? Gibt es einen bestimmten (fachspezifischen oder universellen) Standard, der angewandt werden kann (z. B. ISO 19115 oder Dublin Core)?

d) In welchem Format werden die Metadaten gespeichert und ausgetauscht, in welcher *Syntax* werden sie präsentiert?

e) Welche Metadaten können *automatisch* erhoben werden, wer organisiert die vollständige und korrekte Erfassung der anderen?

f) Welche Voraussetzungen bestehen hard- und softwaretechnisch für die Verarbeitung dieser Metadaten?

g) Welche Vorkenntnisse/Fachkenntnisse sind zum Verständnis / für die Verarbeitung dieser Metadaten erforderlich?

11 Identifikatoren und Informationsobjekte

11.1 Zu identifizierende Informationsobjekte

a) Wie sind die Informationsobjekte definiert und in welchen Verhältnissen stehen sie zueinander? *(z.B.: Werden Daten aus anderen Daten erzeugt oder gibt es logische Beziehungen zwischen ihnen? Sind diese Verhältnisse in einem Informationsmodell oder einer formalen Ontologie dokumentiert?)*

b) Welche Informationsobjekte sind so zentral, dass sie eigene Identifikatoren benötigen?

c) Für welche Informationsobjekte ist es wichtig, dass die Identifikatoren persistent bzw. dauerhaft zitierfähig sind?

11.2 Identifikatoren

a) Was für Identifikatoren werden benutzt (Standards, Syntax etc.)? Sind sie projektübergreifend definiert?

b) Wo werden die Identifikatoren nachgewiesen/aufgelöst? Wird ein externer Resolver-Anbieter in Anspruch genommen?

c) Wer wird die Aktualisierung und Pflege von Identifikatoren während des Projektes / nach Projektende vornehmen?

Kurzfassung:

1. Planung und Erstellung
   - 1.1 Sind alle Rahmendaten des Projekts dokumentiert (Name, Ziele, Finanzgeber, Laufzeit, Partner, Leiter)?
   - 1.2 Wie können bereits existierende Daten integriert/nachgenutzt werden?
   - 1.3 Welche Bedeutung haben die Daten für das Vorhabensziel?
   - 1.4 Wie lassen sich die verwendeten/erzeugten Daten charakterisieren (Datenarten, Formate, Reproduzierbarkeit)?
   - 1.5 Wie werden die Daten erfasst/erstellt?
   - 1.6 Wie groß ist die Datenmenge/Produktionsrate?
   - 1.7 Wie erfolgt die Qualitätssicherung?
2. Auswahl und Aufbewahrungsdauer
   - 2.1 Wieso müssen welche Daten aufbewahrt werden?
   - 2.2 Wann, durch wen und womit erfolgt die Datenauswahl?
   - 2.3 Wie lange müssen die Daten aufbewahrt werden?
   - 2.4 Was geschieht bei Ablauf der Aufbewahrungsdauer?
3. Ingest: Einspeisen und Verantwortungsübernahme
   - 3.1 Wann und wie werden die Daten übergeben/übertragen?
   - 3.2 Wann und von wem werden welche Metadaten erfasst?
   - 3.3 Wie werden Daten und Metadaten validiert?
   - 3.4 Wie wird mit sensiblen Daten umgegangen?
   - 3.5 Sind Rechte und Pflichten von Datenproduzent und -archiv geklärt (Protokollierung, Fehlerbehandlung)?
4. Speicherung und Infrastruktur
   - 4.1 Wer ist während des Projekts und danach für die Speicherung der Daten verantwortlich?
   - 4.2 Mit welchen Technologien und an welchen Orten werden die Daten gespeichert?
   - 4.3 Werden regelmäßig Sicherheitskopien erstellt und überprüft?
   - 4.4 Wie hoch ist die erwartete Datenmenge?
   - 4.5 Gibt es besondere Infrastruktur-Anforderungen für Datentransfer, -zugriff und -nutzung? (Netzwerk-Bandbreite, Hardware etc.)
5. Erhaltungsmaßnahmen
   - 5.1 Sind die eingesetzten Technologien sowie Abhängigkeiten, Nutzungszielgruppe und -anforderungen dokumentiert?

5.2 Wird regelmäßig überprüft, ob sich die Anforderungen, verfügbaren Technologien oder Abhängigkeiten verändert haben?

5.3 Wird die Aufbewahrungswürdigkeit regelmäßig überprüft?

5.4 Gibt es eine Nachfolgeregelung bei einem Wechsel der aufbewahrenden Institution?

6. Zugriff und Nutzung

6.1 Können die Daten auch von Anderen innerhalb oder außerhalb des Projekts genutzt werden?

6.2 Gibt es Verpflichtungen, Daten freizugeben oder nicht freizugeben?

6.3 Welche Einrichtungen/Gruppen werden an den Daten interessiert sein?

6.4 Wie werden Veröffentlichung, Suchbarkeit und Zugriff realisiert?

6.5 Zu welchem Zweck und mit welcher Software werden die Daten voraussichtlich genutzt?

6.6 Wie werden ggf. Erstnutzungsrecht, Nutzungsbeschränkungen oder Lizenzbedingungen durchgesetzt?

6.7 Spielt Interoperabilität eine Rolle?

7. Management, Organisation und Policies

7.1 In welcher Art von Repository werden die Daten aufbewahrt?

7.2 Ist der Bezugsrahmen national oder international?

7.3 Welche Institutionen sind für das Datenmanagement zuständig oder daran beteiligt?

7.4 Haben alle Beteiligten eingewilligt und sind deren Beiträge definiert?

7.5 Ist der Workflow des Datenmanagements beschrieben?

7.6 Sind alle Ressourcen beschrieben und abgeschätzt?

7.7 Welche Anforderungen/Auflagen/Policies müssen berücksichtigt/umgesetzt werden? Von wem?

7.8 Wie wird die Einhaltung der Planung überprüft?

8. Kosten

8.1 Wie hoch sind die Kosten und der Personalaufwand für das Datenmanagement während des Projektes/nach dem Projekt?

8.2 Wer übernimmt die Kosten?

8.3 Stehen alle Verantwortlichen und Beteiligten hinter den Plänen zum Datenmanagement?

9. Rechtliche Aspekte von Forschungsdaten

9.1 Sind die Daten aufgrund des Datenschutzes oder aus anderen Gründen sensibel?

9.2 Werden fremde Daten oder Software verwendet, welche dem Urheberrecht, dem Patentrecht o.ä. unterliegen?

9.3 Unterliegen eigene Daten oder Software dem Urheberrecht/Patentrecht und sind die Nutzungsbedingungen und Rechte geklärt?

9.4 Werden alle notwendigen Maßnahmen zum Datenmanagement eingeräumt?

9.5 Sind evtl. auslaufende Schutzfristen zu berücksichtigen?

10. Metadaten

10.1 Wozu sollen die Metadaten dienen oder benutzt werden?

10.2 Was soll durch Metadaten beschrieben werden?

10.3 Welche Semantik und Syntax wird verwendet?

10.4 Inwieweit können die Metadaten automatisch erstellt werden?

10.5 Welche Voraussetzungen bestehen für die Verarbeitung und das Verständnis der Metadaten?

11. Identifikatoren und Informationsobjekte

11.1 In welchem Verhältnis stehen die Informationsobjekte zueinander?

11.2 Für welche Informationsobjekte werden dauerhafte Identifikatoren benötigt?

11.3 Was für Identifikatoren werden benutzt?

11.4 Wie werden die Identifikatoren nachgewiesen?

11.5 Wer wird die Aktualisierung und Pflege von Identifikatoren vornehmen?