OPENNESS IN FORSCHUNGSPROJEKTEN: PARTHENOS STANDARDIZATION SURVIVAL KIT (SSK)

von Klaus Illmayer

Zusammenfassung: *Die Umsetzung von Open Access und Open Data ist für Horizon 2020 Projekte, die von der Europäischen Kommission gefördert werden, obligat. Überlegungen zu Open Policies und Openness in den Wissenschaften stehen im Mittelpunkt dieses Berichtes, der auf Erfahrungen mit der Erstellung des Standardization Survival Kit (SSK) beruht. Das SSK wurde als ein Tool im Rahmen des Horizon 2020 geförderten Projektes PARTHENOS entwickelt. Daher wird zunächst die Data-Harvesting Plattform von PARTHENOS vorgestellt, um daran die Bedeutung von Openness und Standards zu erläutern. Nach einem Exkurs zu den FAIR Data Prinzipien wird das SSK-Tool beschrieben und wie dieses zu Openness beiträgt. Plädiert wird für eine Dokumentation von Open Workflows und Open Research Methods, wie es mittels dem SSK möglich ist.*

Schlagwörter: *Open Data; Open Science; FAIR Data Prinzipien; Standards; PARTHENOS; Standardization Survival Kit (SSK)*

OPENNESS IN RESEARCH PROJECTS: PARTHENOS STANDARDIZATION SURVIVAL KIT (SSK)

Abstract: *The implementation of Open Access and Open Data is mandatory for Horizon 2020 projects, funded by the European Commission. This report focuses on considerations of Open Policies and Openness in the sciences and humanities, based on experiences gained by creating the Standardization Survival Kit (SSK). The SSK was developed as a tool within the Horizon 2020 funded PARTHENOS project. Therefore, the data harvesting platform of PARTHENOS will be presented first to show the importance of Openness and the application of standards. After a brief discussion on the FAIR data principles, the SSK tool is described and how it contributes to Openness. The SSK can be used for the documentation of Open Workflows and Open Research Methods.*

Keywords: *Open Data; Open Science; FAIR Data Principles; Standards; PARTHENOS; Standardization Survival Kit (SSK)*

DOI: *https://doi.org/10.31263/voebm.v72i2.3221*

Inhalt

1. *Über das Horizon 2020-Projekt PARTHENOS*
2. *Openness in PARTHENOS*
3. *Herausforderungen für Openness*
4. *Warum Standards?*
5. *FAIR Data*
6. *Empfehlung von Standards im Standardization Survival Kit*
7. *SSK und Openness*
8. *Erstellung von Inhalten im SSK*
9. *Open Workflows und Open Research Methods*

1. Über das Horizon 2020-Projekt PARTHENOS

Das 2015 gestartete und von der Europäischen Kommission im Rahmen von *Horizon 2020*[1] auf vier Jahre geförderte Projekt PARTHENOS[2], wird von einem Konsortium aus sechzehn Institutionen getragen. Eine davon ist das *Austrian Centre for Digital Humanities* an der *Österreichischen Akademie der Wissenschaften (ACDH-OEAW)*.

Neben einer Anspielung auf die griechische Mythologie, ist PARTHENOS zugleich ein Akronym, dessen volle Bezeichnung bereits das Projektziel umreißt: „Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies“[3]. Kurz zusammengefasst sollen bereits existierende Daten aus unterschiedlichen Plattformen (den Datengeber_innen) an einem zentralen Ort gesammelt (auch als *Harvesting* bezeichnet), durchsuchbar und dadurch für die Forschung nachnutzbar gemacht werden, wobei zunächst nicht die Inhalte selbst, sondern deren Metadaten unter einem einheitlichen Schema zusammengefasst werden. Die Daten können auf Anforderung von den Datengeber_innen bezogen werden, um in einem „Virtual Research Environment“ (VRE) bearbeitet zu werden. Der Fokus richtet sich auf Kulturerbe-Daten und umfasst institutionell Benutzer_innen aus den Geistes- und Kulturwissenschaften, der Archäologie, dem GLAM-Sektor[4] und den Sozialwissenschaften[5].

Relevant für diesen Beitrag ist, dass für *Horizon 2020* Projekte die Umsetzung von *Open Access* und *Open Data* – zwei wesentliche Aspekte von *Openness* für digitale Forschungsprojekte – als obligatorisch gilt[6]. In Folge

wird insbesondere die Bedeutung von *Open Data* im PARTHENOS Projekt dargelegt. Zugleich wurden im Projekt auch Aktivitäten gesetzt, um *Openness* zu fördern, was am Beispiel der Entwicklung des Tools *Standardization Survival Kit* (SSK)[7] aufgezeigt wird.

2. Openness in PARTHENOS

Es versteht sich von selbst, dass das *Harvesten* von Daten, wenn diese das Kriterium der *Openness* nicht erfüllen, nur bedingt erfolgreich sein kann. Zwar gibt es für Forschungszwecke oft rechtliche Ausnahmeregeln, aber da diese national sehr unterschiedlich ausfallen[8], kann das Bereitstellen von Daten oder zumindest von Metadaten, nur im Rahmen von offenen Lizenzen dauerhaft gewährleistet werden.

Zudem ist eine offene Bereitstellung von Daten förderlich für die Aufrechterhaltung von Forschungsdaten-Lebenszyklen. Entsprechende Anreize zu setzen, wurde deswegen als eine zentrale Aufgabe im PARTHENOS-Projekt festgelegt.[9] Als Forschungsdaten-Lebenszyklen werden die verschiedenen Schritte bezeichnet, die bei der Erstellung von Forschungsdaten erfolgen. So kann ein Lebenszyklus mit der Akquirierung der Rohdaten beginnen (bspw. Digitalisate), darauf folgt die Erstellung eines Schemas, in dem die Rohdaten überführt werden, der Auswertung dieser nun strukturierten Daten und endet schließlich mit der Archivierung. Die archivierten Daten können wiederum für einen neuen Forschungsdaten-Lebenszyklus als Ausgangsbasis verwendet werden.[10]

Sowohl aus Sicht eines *Harvesting* als auch der Umsetzung von Forschungdaten-Lebenszyklen ist die Auseinandersetzung mit *IPR* (Intellectual property right: Geistiges Eigentumsrecht) und *Open Data* sowie *Open Access* von substantieller Bedeutung, da damit eine Rahmensetzung erfolgt, wie weitreichend solche Tätigkeiten umgesetzt werden können. Im PARTHENOS-Projekt wurde dem Rechnung getragen, indem eine Bestandsaufnahme der Rechtssituation auf internationaler Ebene vorgenommen wurde. Zudem wurden Richtlinien ausgearbeitet, um die Frage zu beantworten, welche Daten im Projekt integriert werden können und auf welche Weise die Lizenzierung dieser Daten erfolgt.[11]

Eine zentrale Rolle nehmen für das PARTHENOS-Projekt Repositorien ein, von denen im Regelfall die Metadaten bezogen werden. Eindeutige Informationen zu Benutzungs- und Verwertungslizenzen sind zudem ein Nebeneffekt, wenn auf eine hohe Qualität der dort abgelegten Daten Wert gelegt wird. Im engen Zusammenhang dazu steht eine bewusste Aus-

einandersetzung der Datenproduzent_innen mit Forschungsdatenzyklen, speziell der Identifikation einzelner Forschungsschritte unter Berücksichtigung disziplinenspezifischer Herausforderungen für die Generierung und Beschreibung von Daten.

Warum ist das Zusammenspiel von Forschungsdaten-Lebenszyklen, *Openness*-Kriterien wie Lizenzangaben sowie Repositorien von solcher Bedeutung? Zunächst wird das Gros der Daten für ein *Harvesting* – das Zusammenführen von Daten aus unterschiedlichen Quellen –, wie es PARTHENOS betreibt, von Repositorien bezogen. Einerseits, weil zumeist eine Schnittstelle zum Datenaustausch vorhanden ist und andererseits, weil die Daten dort in dokumentierten Standards abgelegt sind.

Vielen ist vermutlich der Metadaten-Standard Dublin Core (DC)[12] bekannt. Damit stehen zwar nur minimale Informationen zur Verfügung, da diese aber auf einem breiten, gemeinsamen Nenner beruhen, lassen sich Datensätze zumindest ansatzweise miteinander vergleichen und zueinander in Bezug bringen. Eines der bei DC definierten Felder betrifft auch die Lizenz eines Datensatzes[13].

Darüber hinaus gibt es komplexere Datenschemata, die tiefergehende Informationen zur Verfügung stellen. Das hauptsächliche Problem besteht aber darin, dass das Befüllen relevanter Felder über Lizenzierungen nicht überall selbstverständlich ist und oft von der gelebten Praxis in einem Forschungsfeld abhängt. Datenqualität ist somit eine direkte Folge eines Bewusstseins darüber, was mit erzeugten Daten in weiterer Folge passiert bzw. passieren könnte.

Forschungsdatenzyklen verleiten idealerweise zu einem solchen gesamtheitlichen Denken. Nicht nur sollen die Daten für die Auswertungen in einem Forschungsprojekt verwendbar sein, sondern darüber hinaus für Folgeprojekte sowie gänzlich andere Projekte, an die unter Umständen bei der Datenproduktion gar nicht gedacht wurde. Hierbei sind Standardisierungen und das Einhalten von dokumentierten *Workflows* für die Nachnutzbarkeit von Daten von großer Bedeutung. Die Vergabe von offenen Lizenzen rundet ein solches Potential für die Weiterverarbeitung ab und ermöglicht es, Forschungsdaten in einem Zyklus zu halten, bei welchem die Ergebnisse eines Projekts für den Beginn eines neuen Projekts verwendet werden können.

Repositorien sowie die Berücksichtigung von Forschungsdaten-Lebenszyklen sind wichtige Bausteine nicht nur für die Kennzeichnung von Daten als offene Daten, sondern auch für das Überzeugen der Datenproduzent_innen, dass ein solcher Schritt nachhaltig und sinnvoll für die gesamte Forschungsgemeinschaft ist. Projekte wie PARTHENOS wiederum zeigen vor,

wie auf solcher Basis Daten aus anderen Projekten miteinander in Bezug gebracht werden können. Damit wird nicht nur der Nutzen von Daten, die FAIR sind und auf Standards beruhen, beispielhaft vorgeführt, sondern es zeigt sich auch das Potential für die Entwicklung neuer Forschungsfragen.

3. Herausforderungen für Openness

Eine der größten Schwierigkeiten von *Openness* stellt mangelhafte Information über die Lizenzierung von Daten dar. *Open Data* ist als Prinzip weit weniger anerkannt als *Open Access*. Anders gesagt, die Lizenzierung von Publikationen ist weit selbstverständlicher, als dies bei Datensätzen der Fall ist. Im Besonderen gibt es deutliche Unterschiede zwischen Metadaten und Daten. Bei Metadaten wird in der Regel angenommen, dass sie frei zugänglich sind, auch wenn dies in vielen Fällen nicht klar gekennzeichnet ist.

Neben diesen Unklarheiten sind es – speziell in sozialwissenschaftlichen Disziplinen – ethische Überlegungen, die mit manchen der Daten einhergehen. Wenn personenspezifische Informationen in Datensätzen enthalten sind, dann galt bereits vor der Datenschutz-Grundverordnung[14] eine besondere Schutzwürdigkeit, so es sich um lebende Personen handelt. Auch die wohlbekannte Lücke, dass auf Grund der Schutzfristen von 70 Jahren nach dem Tod, in den letzten Jahrzehnten zwar enorme Mengen an Daten produziert wurden, diese aber bei Fehlen von (offenen) Lizenzen aus Urheber_innen- und Verwertungsrechtsgründen nicht oder nur eingeschränkt verwendet werden können, hat zu einem Ungleichgewicht geführt. Ist der Rechtsstatus unklar, bedeutet dies eine zeitaufwendige Recherche, die oft genug zu keinem Ergebnis führt.

Der Anteil an Datensätzen ohne bzw. mit unklaren Lizenzen kann am *Harvesting*-Dienst der europäischen Infrastruktur CLARIN[15], die auf Sprachressourcen spezialisiert ist, aufgezeigt werden. Dies ist auch insofern von Interesse, weil CLARIN Teil des PARTHENOS-Konsortiums ist und Daten für das Projekt beisteuert. Im CLARIN-Suchdienst Virtual Language Observatory (VLO)[16] finden sich mehr als eine Million Ressourcen. VLO gibt eine Übersicht, welche Daten bei unterschiedlichen Datengeber_innen vorhanden sind und wie auf diese zugegriffen werden kann. Damit agiert dieses Service ähnlich wie PARTHENOS, beschränkt sich jedoch auf eine spezifische Domäne.

Ein Blick in die *Availability* der in VLO gelisteten Datensätze zeigt, dass mehr als die Hälfte davon keine Angaben zu Lizenzen beinhalten[17]. In solchen Fällen wird darauf verwiesen, bei den Datengeber_innen nachzufra-

gen. Dies bedeutet nicht, dass keine dieser Datensätze öffentlich zugänglich wären. Das Problem liegt darin, dass die Information darüber schlicht und einfach nicht vorhanden ist. Die Gründe dafür können vielseitig sein: Neben der Nichterfassung kann die Verwendung eines nicht standardisierten Metadatenschemas oder auch technische Unzulänglichkeiten ausschlaggebend sein. Das erschwert eine nachhaltige Arbeit mit solchen Ressourcen, da entweder eine zeitraubende Nachforschung eingegangen oder auf Grund der unsicheren Lizenzinformation auf die Verwendung des Datensatzes verzichtet werden muss. CLARIN VLO wurde hier als Beispiel herangezogen, weil dort die entsprechenden Informationen gut und transparent aufbereitet sind. Was bei vielen anderen Plattformen nur eingeschränkt der Fall ist. Es ist davon auszugehen, dass die Zahl von unklar dargestellten Lizenzen auch bei Daten aus anderen Quellen ähnlich ist.

Die Tragweite dieser mangelnden (Meta)Datenqualität kann am Beispiel von PARTHENOS aufgezeigt werden. Wie bereits geschildert, sollen Daten aus verschiedensten Plattformen so miteinander in Verbindung gebracht werden, dass durch diesen Zusammenschluss neue Erkenntnisse entstehen können. Durch die Verknüpfung von archäologischen mit historischen und sozialwissenschaftlichen Informationen soll es möglich sein, eine bessere Kontextualisierung (z.B. einer Grabungsstätte) herzustellen. Für Forschende der Archäologie können in diesem Fall geschichtswissenschaftliche Daten einen neuen und erkenntnisreichen Blickwinkel bieten.

Wie aber ist vorzugehen, wenn diese Daten prinzipiell zwar vorhanden sind, aber entweder eine unzureichende Lizenzangabe oder gar eine restriktive Benutzungsvereinbarung haben? Im schlechtesten Fall, der aber zugleich für eine technische Lösung meist der pragmatischste ist, werden diese Informationen schlicht und einfach ignoriert. Zwar wäre es möglich, zumindest darauf hinzuweisen, dass Daten vorhanden aber nicht verfügbar sind, wenn aber mehrere Millionen von Daten involviert sind, braucht es dafür einen gut durchdachten Mechanismus. Bei sehr großen Datenmengen können nur automatisierte Verfahren angewandt werden, da eine menschliche Kuration nicht mehr möglich ist.

In PARTHENOS wird ein einheitliches Metadatenschema basierend auf CIDOC CRM[18] eingesetzt. Datengeber_innen durchlaufen einen *Mapping*-Prozess, um ihre Daten auf Basis vorgegebener Regeln zu adaptieren und in Folge auf der PARTHENOS-Plattform anzubieten[19]. Je besser sich die Datenqualität am Ausgangsort darstellt, umso einfacher ist die Integration in das Zielsystem möglich. Falls unklare Informationen bspw. zur Lizenzierung vorliegen, ist eine grundlegende Entscheidung zu treffen, ob solche Daten überhaupt aufgenommen werden sollen. Dabei ist die Lizenzierungs-

frage nur ein Baustein unter mehreren, die es bei einem solchen Vorgehen zu berücksichtigen gilt: Das Verwenden von Standards, die Einhaltung von Richtlinien, die Dokumentation des Datenmodells und die Bereitschaft, die Daten zu teilen, sowohl auf institutioneller als auch technischer Ebene, sind von gleicher Bedeutung für ein funktionierendes Zusammenführen von Daten. Zugleich sind diese Bausteine ein Ausdruck von *Openness*.

Wie kann nun darauf eingewirkt werden, dass ein stärkeres Bewusstsein und Bekenntnis zu den Prinzipien der *Openness* hergestellt wird? In PARTHENOS wird auf eine nachhaltige Disseminationsstrategie und die Unterstützung von Forscher_innen durch *Tools* und *Guidelines* gesetzt. Schließlich gilt es bei den Datenproduzent_innen anzusetzen, um damit in Folge qualitativ hochwertige Daten in der *Harvesting*-Plattform anbieten zu können. Mit dieser Zielsetzung wurden mehrere Initiativen gestartet um ein verstärktes Bewusstsein über die Zusammenhänge zwischen Datenqualität, *Openness* und digitale Forschung herzustellen: eine Trainingswebsite mit Übungsmodulen[20], ein interaktiver *Hub* mit relevanten Publikationen[21], eine Zusammenstellung von Richtlinien aus den beteiligten Disziplinen (*Policy Wizard*)[22] und das bereits erwähnte *Standardization Survival Kit* (SSK), das in der Folge detailliert vorgestellt wird.

4. Warum Standards?

Standardisierung mag für einzelne Projekte oft als eine Belastung gelten, da es eine Tendenz dazu gibt, auf eigene Erfahrungen und Vorstellungen zu setzen, statt Vorgaben einzuhalten, die durch Standards gemacht werden. Diese werden oft als Einschränkungen wahrgenommen, wobei übersehen wird, dass ein Standard viele Blickwinkel berücksichtigt, die unter Umständen zunächst wenig bis gar nichts mit einem Einzelprojekt zu tun haben können. Nichtsdestotrotz lohnt es sich mittel- und langfristig, auf *Community*-Standards zu setzen. Insbesondere im Hinblick auf die vorgestellten Überlegungen zum Forschungsdaten-Lebenszyklus oder auch für die Teilnahme an projektübergreifenden Infrastrukturen wie PARTHENOS oder CLARIN. Oft lässt sich erst durch das Einhalten von gemeinsamen Vorgaben das Potential von Standards erkennen.

Leider ermöglicht die Genese vieler Forschungsprojekte keine solche Erfahrung, da nach dem Ende der Finanzierung oft genug auch die Arbeit an den Daten endet. Falls nun nicht auf Standards gesetzt wurde, ist es für Folgeprojekte schwierig bis unmöglich, die Weiterverarbeitung der erzeugten Daten aufzunehmen. Es handelt sich dabei um ein struk-

turelles Problem, dem mit Erfahrungsaustausch entgegengewirkt werden kann.

Häufig ist aber ein viel grundlegenderes Problem zu berücksichtigen, nämlich schlicht das mangelnde Wissen über die Verwendung von Standards in einer *Community* bzw. das Fehlen solcher gemeinsamen Standards. Letzteres gilt es aufzuzeigen und darauf zu insistieren, dass eine gemeinsame Anstrengung unternommen wird, diese Lücke zu schließen. Für den häufigen Fall, dass Standards nicht bekannt sind, sollte relativ einfach Abhilfe geschaffen werden können.

Dies ist das erklärte Ziel des SSK, das als Tool zum einen auf die Verwendung von Standards in einer *Community* hinweist – und damit eine *standards literacy* (Kompetenz für Standards) aufbaut – und zum anderen dazu motivieren soll, solche Standards auch einzusetzen[23]. Der grundlegende Gedanke für die Entwicklung des SSK ist, anhand erfolgreicher Projekte die Verwendung und den Nutzen von Standards aufzuzeigen. Davon profitiert auch der *Openness*-Gedanke, schließlich demonstrieren die vorgestellten Projekte, das *Open Data* für übergreifende Initiativen eine wichtige Voraussetzung ist. Die Berücksichtigung von Standards und der damit verbundenen Erstellung von strukturierten Daten führt zu den *FAIR Data*-Prinzipien.

5. FAIR Data

Unter FAIR werden die Begriffe *Findable*, *Accessible*, *Interoperable* und *Re-Useable* zusammengefasst[24]. Zwar legt *Accessible* ein *Openness* nahe, fordert dies aber nicht explizit ein, da aus oben genannten Gründen dies nicht immer gewünscht bzw. rechtlich möglich ist[25]. Es gilt hingegen die Regel: Mache deine Daten FAIR und versuche zugleich, sie auch *Open* zu halten. Mag dies zunächst als Einschränkung empfunden werden, so ist dies doch ein pragmatischer Umgang mit ethischen und rechtlichen Rahmenbedingungen, die für manche Daten von Bedeutung sind. Dies gilt insbesondere für ein internationales Daten-*Harvesting*, wie im PARTHENOS-Projekt.

Eine Stärke der FAIR-Prinzipien ist die Berücksichtigung formaler Kriterien für die Generierung von Daten. Damit werden Maschinen-zu-Maschinen-Verfahren unterstützt, die in weiterer Folge einen einfach zu skalierenden Datenaustausch anregen. Insbesondere der *Interoperable*-Ansatz ist dabei von großer Bedeutung. Hier wird zudem eine Lücke im *Open Data*-Zugang geschlossen. *Openness* bringt nämlich nur wenig, wenn die Daten schwer oder gar nicht zu interpretieren sind. Wird auf

ein selbstentwickeltes Datenformat gesetzt, das – was in solchen Situationen häufig der Fall ist – über keine ausreichende Dokumentation verfügt, besteht die Gefahr, dass die Daten für eine weitere Verwendung unbrauchbar sind. Die Anwendung von FAIR auf Daten verpflichtet dazu, diesen Zusammenhang bei der Datenerstellung verstärkt in Betracht zu ziehen.

Accessibility und *Findability* wiederum zielen auf die eindeutige Identifikation von Daten ab z.B. vermittels *Identifier*. Auch hier ist der Gedanke des Austausches zentral. Wenn aufbauend auf diesen strukturellen Überlegungen und Adaptionen die Daten zudem noch *Open* lizenziert werden, kann von einer gelungenen und nachhaltigen Datenerstellung gesprochen werden, die nicht nur für mehr Sichtbarkeit, sondern auch für die *Re-Usability* der Daten sorgt. FAIR und *Open* schließen sich nicht aus, vielmehr ergänzen sie sich: Daten, die nur *Open* sind, sind eventuell auf struktureller Ebene unbrauchbar, wohingegen Daten, die FAIR sind, ideal für *Openness* adaptierbar sind und erst damit Relevanz für andere Projekte erlangen.

6. Empfehlung von Standards im Standardization Survival Kit

Wie trägt nun der SSK dazu bei, Überlegungen zu *Openness* im Bereich der Anwendung von Standards aufzugreifen und zu kommunizieren? Bisherige Plattformen, die sich Standards widmen, tendieren dazu, eine Liste aller möglichen Formate, Ontologien, usw. zu erstellen[26]. Das ist von Bedeutung, wenn bereits ein Bekenntnis zu Standards erfolgt ist und es erste Anhaltspunkte gibt, welche Standards eingesetzt werden können. Der SSK geht einen anderen Weg, da es die Lücke zu schließen gilt, zu denen, die nicht über dieses Vorwissen verfügen.

Ausgangspunkt sind sogenannte Szenarien, die einen Arbeitsablauf – oft auf Basis eines Forschungsdaten-Lebenszyklus – abbilden, wobei erfolgreiche Forschungsprojekte die Grundlage dafür liefern. Aktuell umfasst das SSK an die dreißig solcher Szenarien, aus verschiedenen Disziplinen. Zum Beispiel aus dem Bereich der Linguistik ein Szenario zur Erstellung eines digitalen Wörterbuches, aus der Literaturwissenschaft eines, wo die Extraktion von Text aus Bildern/Digitalisaten beschrieben wird oder aus der Archäologie eines zum Einsatz von Laser für die Restauration von Objekten[27]. Die Szenarien sind unterteilt in Teilschritte, die eine ideale Vorgehensweise im Detail beschreiben. Daran angeschlossen sind Empfehlungen, welche Standards eingesetzt werden sollen und Hinweise auf

weiterführende Anleitungen, Dokumentationen oder Forschungspapiere. Anhand der Szenarien und ihrer Teilschritte wird der Nutzen von Standards aufgezeigt und zugleich darüber informiert, welche Standards in den verschiedenen Forschungsdisziplinen angewandt werden.

Der Einstieg in den SSK ist niederschwellig angelegt. Aus einer Liste kann zunächst eine Forschungsdisziplin und dann jenes Szenario ausgewählt werden, das am besten zur eigenen Methodik passt. Durch eine kontinuierliche Erweiterung der Inhalte werden mehr und mehr Fälle abgedeckt, um damit die Bedeutung von Standards möglichst breit zu vermitteln.

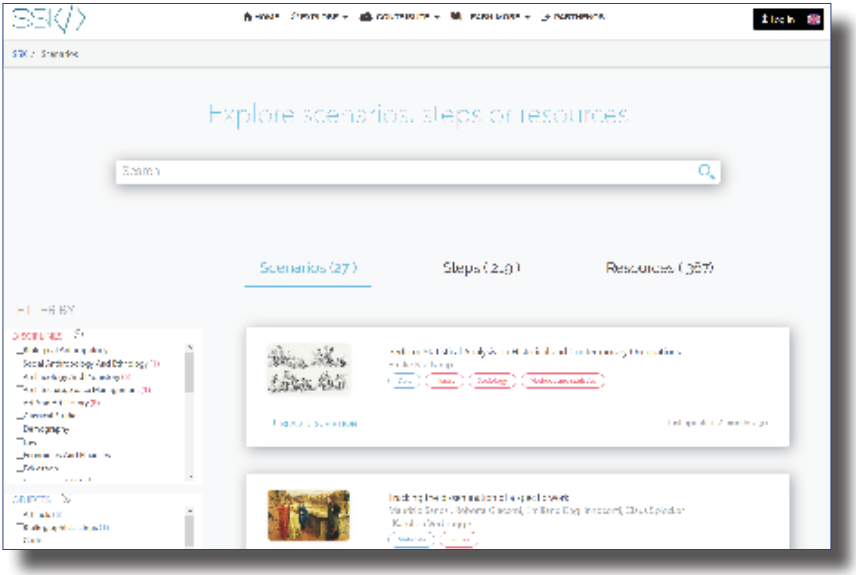

Abb. 1: Standardization Survival Kit (SSK) – Übersichtsseite (http://ssk.huma-num.fr/#/scenarios)

7. SSK und Openness

Zum einen wird in den Szenarien des SSK auf die Vorteile von *Open Data* hingewiesen, um damit eine Argumentation für die Anwendung von *Openness* in Projekten zu geben. Zum anderen ist es die Methodik des SSK selbst, die von *Openness* geprägt ist. Denn es ist die erklärte Absicht, transparent darzulegen, wie digitale Verfahren funktionieren und implementiert werden. Zudem ist das SSK selbst offen aufgebaut. Der Sourcecode liegt auf GitHub, ebenso die Ausgangsdaten, die in die Website

importiert werden[28]. Die Ressourcen werden über das Bibliografie-Tool Zotero verwaltet[29]. Verlinkt wird zudem auf bereits etablierte Vokabularien wie TaDiRAH[30]. Zugleich wird für Detailinformationen über Standards auf Datenbanken wie das CLARIN *Standards Information System* verwiesen.

Bereits bei dieser Auflistung zeigt sich, wie ein gelungenes Zusammenspiel mehrerer Komponenten und Plattformen hergestellt werden kann, wenn auf *Open Data* gesetzt wird. Das betrifft auch die Daten des SSK, die die Szenarien und Teilschritte beschreiben. Diese entsprechen ebenso den FAIR Data- und *Open Data*-Prinzipien und liegen strukturiert im XML-Format der *Text Encoding Initiative* (TEI)[31] vor.

8. Erstellung von Inhalten im SSK

Für Informationen zu den Workflows und den eingesetzten Standards, braucht es Expert_innen, die befragt und deren Wissen in das Datenmodell des SSK übersetzt werden. Somit ist das Erstellen eines Szenarios mitunter ein aufwendiger Akt, auch weil ein reflektierter Prozess nötig ist, um über das eigene Tun auf formaler Ebene zu sprechen.

Gleichzeitig besteht der Anspruch, möglichst viele Praxen einer Forschungs-*Community* zu repräsentieren. Insofern sind auch alternative Perspektiven zu berücksichtigen. Als zielführend haben sich dafür Workshops erwiesen, bei denen mehrere Expert_innen ihre Zugänge zur Diskussion stellen und entweder ihr eigenes oder ein gemeinsames generisches Szenario zusammenstellen[32].

Aktuell werden diese Informationen aufgezeichnet und von einer Redaktion in das Datenmodell umgewandelt. Es gibt Pläne, dies noch stärker zu formalisieren und entsprechende Eingabemasken zur Verfügung zu stellen, so dass noch einfacher Szenarien in das SSK eingearbeitet werden können. An einer Umsetzung wird derzeit gearbeitet.

Als entscheidender Vorteil für den Informationsgehalt im SSK erweist sich der offene Zugang bei der Erstellung von Szenarien und deren Teilschritte. Indem Abläufe transparent dargelegt und auch Schwierigkeiten angesprochen werden – so verbirgt sich hinter dem Anschein ausgeklügelter Techniken oftmals bloß eine einfache manuelle Tätigkeit –, bietet das SSK einen offenen Einblick in die Herausforderungen digitaler Forschung. Wichtig ist dabei – speziell auch für spätere Auswertungen –, dass diese formalisierte Erhebung von Abläufen in Form von strukturierten Daten erfolgt, die FAIR und *Open* sind.

9. Open Workflows und Open Research Methods

Im Hinblick auf *Openness* kann aus dem Ansatz und den Erfahrungen des SSK abgeleitet werden, dass viele Komponenten erforderlich sind, die erst im Zusammenwirken offene Strukturen erzeugen bzw. verstärken. Auch wenn *Open Access* vermutlich die größte Aufmerksamkeit im öffentlichen Diskurs einnimmt, sind Erweiterungen im Feld der *Openness* – wie etwa *Open Data* – eng damit verzahnt. Für den generellen Aufbau von offenen Strukturen zeigt sich im Zusammenwirken deren Wichtigkeit. Durch die zunehmende Vernetzung von Daten, dem vermehrten Aufbau von gemeinsamen Plattformen und neuen technischen Möglichkeiten – insbesondere im Bereich *Linked Open Data* – ergeben sich dadurch neue Betätigungsfelder.

Für die Forschung mit digitalen Methoden bedeutet die Offenlegung methodischer Zugänge eine nicht zu unterschätzende Herausforderung. Vor allem die Erstellung strukturierter Daten, die die methodischen Zugänge formalisiert beschreiben sollen, erweist sich nicht selten als schwierig. Während die Erläuterung der angewandten Forschungsmethode in Publikationen gang und gäbe ist, kann dies für deren Abbildung auf Datenebene noch nicht festgestellt werden. Der Forschungsdaten-Lebenszyklus gibt dafür zumindest einen ersten formalen Rahmen ab.

Mag eine solche digitale Formalisierung in Bereichen der Technik- und Naturwissenschaft bereits üblich sein, so ist dies für die Geistes- und Kulturwissenschaften seltener der Fall, auch weil sich manches – speziell wenn es um kreative Zugänge geht – nicht ohne weiteres formalisieren lässt. Aber auch diese Tatsache lässt sich beschreiben und kontextualisieren.

Gewonnen wird mit der Dokumentation und Veröffentlichung von *Open Workflows* und *Open Research Methods* nicht nur ein Einblick in Forschungsprozesse, sondern auch ein Austausch inner- und außerhalb der Forschungs-*Community*. Zudem werden dem *Openness*-Baukasten damit weitere grundlegende Werkzeuge hinzugefügt.

Das SSK ist dafür ein hilfreiches Tool. Anzumerken ist allerdings, dass noch mehr Erfahrungen zur Tragfähigkeit der Abbildung von Forschungsprozessen sowie der Auseinandersetzung mit Bedenken, die vor allem auf die mögliche Banalisierung komplexer Verfahren abzielen, erforderlich sind. Dennoch kann bereits jetzt festgehalten werden, dass für eine Bewusstseinsbildung hinsichtlich der Wichtigkeit von Standards sowie von FAIR und *Open Data* mit dem SSK ein hilfreiches Instrument zur Verfügung steht. Darüber hinaus sollten Plattformen und Projekte wie PARTHENOS sowie große und kleine digitale Infrastrukturen ebenso proaktiv auf diese Aspekte hinweisen.

Für die nachhaltige Fortführung des SSK wurde durch die Übernahme der Redaktionstätigkeit in die Arbeitsgruppe *Guidelines and Standards* der europäischen Infrastruktur DARIAH[33] ebenfalls gesorgt. Hier befindet sich auch der Kontaktpunkt für Rückmeldungen, einer Mitwirkung am SSK oder Anregungen für neue Szenarien, zu dem an dieser Stelle herzlich eingeladen wird[34].

Mag. Dr. Klaus Illmayer
ORCID: https://orcid.org/0000-0001-7253-996X
Österreichische Akademie der Wissenschaften,
Austrian Centre for Digital Humanities (ACDH-OEAW)
E-Mail: klaus.illmayer@oeaw.ac.at

Literatur

Drude, Sebastian; Di Giorgio, Sara; Ronzino, Paola; Links, Petra; van Nispen, Annelies; Verbrugge, Karolien; Degl'Innocenti, Emiliano; Oltersdorf, Jenny; Stiller, Juliane; Spiecker, Claus (2016): PARTHENOS D2.1 Report on User Requirements. Zenodo. https://doi.org/10.5281/zenodo.2204561

Puhl, Johanna; Andorfer, Peter; Höckendorff, Mareike; Schmunk, Stefan; Stiller, Juliane; Thoden, Klaus (2015): Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften. (DARIAH-DE Working Papers 11). Göttingen: DARIAH-DE. http://resolver.sub.uni-goettingen.de/purl/?dariah-2015-4

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, Ijsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie, ... Mons, Barend (2016): The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

* Der letzte Zugriff auf alle angegebenen Links erfolgte am 31.10.2019.

1 Dabei handelt es sich um ein EU-Rahmenprogramm, das auf die Förderung von Forschung und Innovation in der Europäischen Union abzielt. Es setzt die Teilnahme von mehreren Institutionen aus mehreren Ländern der EU und assoziierter Staaten voraus, vgl. European Commission: Horizon 2020, https://ec.europa.eu/programmes/horizon2020/

2 PARTHENOS Project, http://www.parthenos-project.eu/. Das Grant Agreement der Europäischen Kommission hat die Nummer 654119,

siehe auch: CORDIS: Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies, https://cordis.europa.eu/project/rcn/194932/factsheet/en

3 Siehe Abschnitt "Who or what is PARTHENOS" in, PARTHENOS: FAQs, http://www.parthenos-project.eu/about-the-project-2/faq

4 GLAM steht für Galerien (Galleries), Bibliotheken (Libraries), Archive (Archives) und Museen (Museums).

5 Zu den anvisierten Benutzer_innengruppen vgl. Sebastian Drude et al: PARTHENOS D2.1 Report on User Requirements, 2016, https://doi.org/10.5281/zenodo.2204561, S. 11–16.

6 Vgl. Den Abschnitt "Open access & Data management" in, European Commission: Horizon 2020 Online Manual, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm, wobei sowohl "open access to scientific publications" als auch "open access to research data" explizit erwähnt wird.

7 PARTHENOS: Standardization Survival Kit, http://ssk.huma-num.fr/

8 Vgl. Sebastian Drude et al: PARTHENOS D2.1 Report on User Requirements, 2016, https://doi.org/10.5281/zenodo.2204561, S. 22ff.

9 Diese Tätigkeit war im Fokus des PARTHENOS *Work Package* 3: "Common policies and implementation strategies", dessen Ergebnisse u.a. als *Deliverables* verfügbar sind, vgl. PARTHENOS: Project's deliverables, http://www.parthenos-project.eu/resources/projects-deliverables#1523355493967-5f00dfe2-070e

10 Für weiterführende Informationen siehe Johanna Puhl et al: „Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften", Göttingen: GEODOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2015 (DARIAH-DE working papers 11), http://resolver.sub.uni-goettingen.de/purl/?dariah-2015-4

11 Vgl. Sebastian Drude et al: PARTHENOS D2.1 Report on User Requirements, 2016, https://doi.org/10.5281/zenodo.2204561, S. 73ff.

12 Dublin Core Metadata Initiative: DCMI Metadata Terms, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

13 Es handelt sich dabei um das Element *dc:rights*.

14 Vgl. auch die aufgezählten Gesetze und Verordnungen auf der Website der österreichischen Datenschutzbehörde, Österreichische Datenschutzbehörde: Datenschutzrecht in Österreich, https://www.dsb.gv.at/gesetze-in-osterreich

15 CLARIN steht für "Common Language Resources and Technology Infrastructure" und ist als ERIC (European Research Infrastructure Con-

sortium) anerkannt, CLARIN ERIC: CLARIN – European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/

16 CLARIN ERIC: Virtual Language Observatory, https://vlo.clarin.eu/

17 Stand vom 4.7.2019: Es können in VLO alle Ressourcen betrachtet werden („See all records"), um danach den Aspekt „Availability" durch Anklicken zu öffnen und eine Übersicht über die Verfügbarkeit dieser Ressourcen zu erhalten.

18 CRM steht für „Conceptual Reference Model", während CIDOC das Komitee für Dokumentation der internationalen Museumsorganisation ICOM ist, vgl. http://cidoc-crm.org/

19 Die Plattform kann über das *Virtual Research Environment* eingesehen werden, PARTHENOS: Virtual Research Environment, https://parthenos.d4science.org/web/parthenos_vre; zentraler Bestandteil ist die Discovery-Plattform, PARTHENOS: PARTHENOS Discovery, https://parthenos.acdh.oeaw.ac.at/

20 PARTHENOS Training: https://training.parthenos-project.eu/

21 PARTHENOS Hub: http://www.parthenos-project.eu/portal/the-hub

22 PARTHENOS Policy Wizard: http://www.parthenos-project.eu/portal/wizard/policy-wizard

23 Begleitend dazu wurde im Projekt PARTHENOS eine Broschüre entwickelt, die darüber informiert, warum Standards eingesetzt werden sollen, PARTHENOS: Why standards?, http://www.parthenos-project.eu/Download/Flyer-Parthenos_standards_ls.pdf

24 Initiiert 2016 durch einen Artikel in *Scientific Data*: Mark D. Wilkinson et al: *The FAIR Guiding Principles for scientific data management and stewardship*, in: Scientific Data 3, 2016, https://doi.org/10.1038/sdata.2016.18; aktuell existieren mehrere Initiativen für die Verbreitung und Umsetzung der FAIR-Prinzipien z.B. eine FORCE11-Gruppe „The Fair Data Principles", https://www.force11.org/group/fairgroup/fairprinciples oder auch GO FAIR, https://www.go-fair.org/

25 Siehe auch GO FAIR: What is the difference between "FAIR data" and "Open data" if there is one?, https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/

26 Bspw. CLARIN-D: CLARIN Standards Information System, https://clarin.ids-mannheim.de/standards/

27 Alle diese Beispiele lassen sich auf der Szenario-Übersichtseite des SSK finden: http://ssk.huma-num.fr/#/scenarios. Aktuell sind die Szenarien und die Teilschritte nur auf Englisch verfügbar, die Übersetzung von Szenarien ist geplant.

28 PARTHENOS: Development of the Standardization Survival Kit, https://github.com/ParthenosWP4/SSK. Der Code und die Daten sind lizenziert unter der *Creative Commons*-Lizenz CC BY 4.0.
29 PARTHENOS: SSK-PARTHENOS, https://www.zotero.org/groups/427927/ssk-parthenos
30 Taxonomy of Digital Research Activities in the Humanities (TaDiRAH), https://github.com/dhtaxonomy/TaDiRAH
31 Text Encoding Initiative (TEI): https://tei-c.org/
32 Zuletzt fanden Workshops im Februar 2019 in Wien statt, zu denen Expert_innen aus der Linguistik eingeladen waren, sowie im Mai 2019 in Tours mit Expert_innen aus der Archäologie.
33 DARIAH-EU: WG Guidelines and Standards, https://www.dariah.eu/activities/working-groups/guidelines-and-standards/
34 Kontaktaufnahme per E-Mail an gist@dariah.eu

Funding

PARTHENOS wurde im Rahmen des Forschungs- und Innovationsprogramms der EU Horizon 2020 (2014-2020) und des Calls "H2020 - EU.1.4.4.1.1.1 – Entwicklung neuer Forschungsinfrastrukturen von Weltrang" unter der Grant Agreement Nr. 654119 gefördert.