Fragebogen: Kommentare

Hauke Sonnenberg

28 September 2017

Es folgen die Kommentare, die von den Workshopteilnehmern zu den Fragen im Fragebogen bezüglich ihrer aktuellen Einschätzung (IST-Zustand) und bezüglich ihrer Wunschvorstellungen (SOLL-Zustand) abgegeben wurden.

Daten finden, IST-Zustand

- Mal so, mal so, je nachdem, wie geordnet es im Projektordner aussieht. Bei der Windows-Suche weiß man zumindest, dass alles durchsucht wurde, was manuell oft nicht möglich ist.
- meist nutze ich die Suche, das geht am schnellsten, wenn ich ungefähr weiss wie die Datei heisst die ich suche
- Wenn ich Daten von anderen suche.
- Daten aus fremden Projekten kann ich kaum selbstständig finden. Daten, die ich selbst abgelegt habe, kann ich sehr gut finden
- Ordnerstruktur auf Toplevel noch eindeutig, darunter individuell und daher nicht mehr nachvollziehbar. Viele Doppelungen von Dateien (z.B. Sicherheitskopien, Arbeitsstand), aktuelle Version der Datei daher nicht eindeutig zu identifizieren. Oft nur mit Mühe oder durch Nachfragen Daten auffindbar, gerade für ältere Projekte.
- Oftmals habe ich die Daten selbst abgelegt. Wenn ich nach alten Daten suche brauche ich meist zwei Versuche.
- Ich suche weniger oft Rohdaten als ausgewertete Daten und Präsentationen mit Projektergebnissen. Insbesondere ist mir oft nicht klar, welches die letzte Version der Labordaten ist, daher frage ich i.d.R. den zuständigen Mitarbeiter. Als problematisch sehe ich die Unterordner von Studierenden, da so neben dem Datenordner in “Data_Workpackage” weitere Ordner mit Daten und Auswertungen vorliegen.
- Für mich ist die Baumstruktur und die Daten die in den Unterordnern liegen nicht immer logisch. Ist aber auch sicherlich von Projekt zu Projekt bzw. von Projektleiter und Projektmitarbeiter unterschiedlich.

Daten finden, SOLL-Zustand

- Gut wäre eine immer aktuelle Übersicht über vorhandene Daten in Bezug auf Art (Messgröße), Herkunft (z.B. BWB, Senat), Umfang (Anzahl Dateien, Speicherplatz), Ort der Messung, Zeitintervall. Im idealen Fall könnte man gezielt nach diesen Gesichtspunkten suchen.
- es ist fraglich ob das wirklich möglich ist
- wichtig ist, mich auch in “fremden” Projekten zurecht zu finden
- Bei Ende der Projekte sollte ein Aufräumen der Projektdaten Pflicht sein. Aktuellste Dateien kenntlich machen, Daten zu Berichten zuordnen, unnötige Doppelungen löschen!!
- Jeder funktioniert intuitiv anders, daher ist dieser Zustand eher utopisch
- Für mich ist eine klarere Struktur aus der die Aktualität der Daten hervorgeht besonders wichtig.
- Ich wünsche mir eine gute logische Ordnerstruktur, die ich ohne großen Aufwand auch anwenden n bzw. wo ich ohne großen Aufwand auch schnell Informationen (administrative bzw. auch Daten aus Versuchen) finden kann, um somit auch effizient arbeiten zu können.

Daten bekommen, IST-Zustand

- Daten im Anhang von E-Mails lege ich zusammen mit einem PDF-Ausdruck der E-Mail ab. Das erfordert Disziplin und das mache ich nicht konsequent und nicht immer einheitlich. Es ist schwierig, sich neue Ordner-/Dateinamen auszudenken. Es fehlt eine verbindliche Vorgabe.
- einen Schreibschutz erstelle ich jedoch nicht
- Rohdaten lege ich auf dem Netzwerk ab, Emails werden in Outlook einsortiert und sind so immer auffindbar. Rohdaten werden teilweise bearbeitet, kein Schreibschutz oder Metadatei. Eigene Berechnungen werden teilweise direkt in den Daten durchgeführt (Excel).
- Wenn ich Rohdaten von BWB erhalte, leite ich diese an den zuständigen Mitarbeiter weiter.
- Ich lege die Daten in einem Unterordner im entsprechendem Unterordner von "Data-Work-Packages" ab, aber ohne eine standarisierte Ordner- bzw. Dateinomenklatur
- Falls ich wichtige Daten bekommen, belasse ich Sie im Posteingang bzw. leite diese an die entsprechenden Projektmitarbeiter für weiteres Datenmanagment weiter.

Daten bekommen, SOLL-Zustand

- Es sollte festgelegt sein, welche und an welchem Ort Informationen über Dateien (insbes. von Projektpartnern) gespeichert werden sollten.
- das wäre sehr sinnvoll
- Daten sollen so abgelegt werden, dass sie nicht verändert werden können oder verloren gehen.
- Rohdaten sollten abgespeichert werden, aber die Originalversion ist im Outlook immer nachvollziehbar. Schreibschutz halte ich für nicht notwendig bei meinen Daten.
- Die Metadatehn muss man erstmal haben.
- Ich finde diese Vorgangweise im Moment für mich in Ordnung

Daten verifizieren, IST-Zustand

- Da ich oft mit großen Datenmengen aus einer Vielzahl von (CSV-)Dateien zu tun habe nutze ich meist selbst geschriebene R-Skripte. Bei kleinen Datenmengen (z.B. die noch auf einen Bildschirm passen, würde vermutlich nur eine Überprüfung anhand einfacher Datenqualitätskriterien durchführen
- Oft werden Prüfungen erst gemacht, nachdem Unplausibilitäten aufgetreten sind.
- Daten werden auf Vollständigkeit und dann auf Plausibiltät geprüft.
- Bilanzdaten prüfe ich generell auf Plausibilität durch mehrere Schritte, u.a. Vergleich mit anderen Projekten/Prozessen, Übertragung ins Modell oderSchließung der Massenbilanz. Oft werden die Daten noch mal von mir aufbereitet und zurück an den Lieferanten geschickt für einen Querabgleich, ob alle Daten sauber angekommen sind.
- Ich bin mir dessen, bewusst, dass das Vorgehen suboptimal ist. Oft bekommt man aber auch auf Nachfrage keine zufriedenstellenden Informationen bzw. hat wenig Zeit zur Datenprüfung. Ich wurde gerade angerufen, was auf der KA WEstwitz zu beachten ist, da sich unsere Partner dort mal wieder spontan eine Reparatur überlegt haben... also so viel zum Thema Zeit...
- Zur Datenprüfung wende ich i.d.R. einfache Darstellungen der Daten in Excel an. Die Validiierung erfolgt auf Basis von Erfahrungswerten (z.B. typische Konzentrationbereich von Spurenstoffen, typische Entfernungsraten oder typische Ablaufwerte einer Kläranlage), bei Bedarf erfolgt dann validierung der Rohdaten anhand der Laborbücher z.B. des BWB-Labors.
- In meinem Fall spreche ich hier von z.B.: SAP Daten. Daten aus Versuchen, die mir geschickt werden, werden von mir nur kurz geprüft (einfache Datenqualitätskriterien) und dann an die Projektmitarbeiter weitergeleitet mit dem Vertrauen, dass diese die Daten dann auch kritisch prüfen.

Daten verifizieren, SOLL-Zustand

- Gut wäre eine vorgegebene Checkliste von Mindest- und optionalen Prüfungen. Wir sollten auf existierende Methoden zurückgreifen können. Dazu sollten die zu plausibilisierenden Daten in einem standardisierten Format (z.B. csv, am besten mit standardisierten Spaltennamen) vorliegen.
- Vielleicht wären Mindestanforderungen für versch. Datenquellen sinnvoll, z.B. Laboranalysen sollten enthalten: Messmethode, Nachweisgrenze, etc., Geländemessungen sollten enthalten: mit welchem Gerät wurde gemessen, Messdatum, etc.
- eine Handlungsempfehlung ist auf jeden Fall hilfreich
- Daten der Projektpartner sind grundsätzlich vertraulich, daher können wir Rohdaten nicht freigeben. Originalquellen sind für mich nachvollziehbar (z.B. über Email), aber nicht für andere, das könnte man ggf. verbessern. Plausibilitätsprüfung gehört immer zum Qualitätscheck, das läuft sowieso bereits ab und ist sehr fallspezifisch.
- Ich selbst werde immer seltener mit Rohdaten umgehen. Für Labordaten könnte z.B. eine automatisierte Einordnung der Daten erfolgen, bei denen die Daten mit denen der Vergangenheit abgegelichen werden ("Erwartungsintervall" oder "Balanced-Score-Card" (siehe BWB-LIMS))

Daten säubern, IST-Zustand

- Da ich oft mit großen Datenmengen aus einer Vielzahl von Dateien zu tun habe nutze ich selbst programmierte R-Skripte.
- Vgl. Verifizierung. Ein manuelles Vorgehen muss nicht schlechter sein als ein automatisiertes Vorgehen.
- Bei sehr großen Datenmengen, die direkt über Schnittstellen von einem Server abgegriffen werden, ist eine Bereinigung schwierig auszuführen
- Bei einer überschaubaren Anzahl an Daten ist die manuelle Bereinigung schneller. Bei Online-Daten sollte mit Skripten gearbeitet werden.
- Meine Daten sind meist nicht umfangreich (keine Listen von Messwerten) und daher nicht automatisch zu bereinigen. Übertragungsfehler werden manuell geprüft und ausgebessert, z.B. über Querabgleich zu anderen Projekten oder Rückfrage zum Lieferanten.
- Ich verwende R nicht da mein Bewusstsein, sondern die Datenmengen so groß sind und ich nicht wüsste, wie es in Excel funktioniert
- Lästige Sache. . . .Hat mich in meiner Doktorarbeit Tage gekostet.
- Die Datenbereinigung wird von mir manuell in R vorgenommen. Formatierungsprobleme versuche ich in der Regel durch die Importfunktionen von R zu beseitigen. Bis jetzt ist eine visuelle Untersuchung der Daten ausreichend gewesen, um z.B. outliers, unplausible Werte, Zeitstempelprobleme, u.a., zu finden.

Daten säubern, SOLL-Zustand

- Sollte zusammen mit Verifizierung diskutiert werden. Wichtig ist, dass die Änderungen dokumentiert sind und nachvollzogen werden können. Existierende Methoden sollten wiederverwendet werden können. Dazu ist eine gute Dokumentation, am besten mit Beispielen, erforderlich.
- Bei geringer Datenanzahl müssen keine programmierten Skripte verwendet werden. Eine nachvollziehbare Datenbereinigung muss jedoch angewendet werden.
- Datenformate lassen automatische Säuberung nicht zu, müssen manuell nach Erfahrungswerten und Plausibilität geprüft und gesäubert werden.
- egal in welche Sprache. es muss definiert und dokumentiert werden, wie die Daten bereinigt wurden und warum. Es hängt von der Aufgabe und Ziele ab.

Daten analysieren, IST-Zustand

- Zu diesem Punkt könnte ich alle 4 Antworten ankreuzen, ich nutze auch Excel (schnell und einfach), das ist auch okay sofern ich mir der Risiken bewusst bin
- Die Art des Datensatzes bestimmt das Vorgehen
- Für wenig Daten ist excel sinnvoll, bei großen Datenmengen reicht excel nicht mehr.
- Hohe Datenmenge wird vorab über Pivottabellen gezielt reduziert und dann analysiert. Keine statistische Auswertung der Ergebnisse notwendig.
- Mehr brauche ich nicht. Weitergehende Analysen sind nicht mein Schwerpunkt.

Daten analysieren, SOLL-Zustand

- Je komplizierter die Analyse, desto eher sollte eine Skriptsprache gegenüber Excel favorisiert werden. Gut wäre ein Überblick darüber, welche Methoden von wem zu welchem Zweck in welchen Projekten angewendet wurden. So könnten wir besser auf existierendem Wissen aufsetzen.
- Auf Grund der großen Datenmengen muss eine Software genutzt werden.
- Pivottabellen liefern viel Flexibilität und sind einfach zu handhaben. Datenformate brauchen keine statistische Aufbereitung hinsichtlich Korrelationen oder ähnliches.
- 1. Stringentere Verwendung des Begriffs “signifikante” Veränderung inkl. der vorherigen statitischen Analyse

2. R ist für viele Labordaten (geringer Anzahl Messwerte meist unter 50) evt. zu hoch gegriffen.
3. Statt ein theoretischen Schulung zu Statistik sollten ggf. einzelnen MA zu spezifischen Fragestellungen und deren Umsetzung z.B. in Origin geschult werden.
4. Datenanalyse sollte sich stärker an der Fragestellung und der Art der Daten resp. der Datenmenge ausrichten, evt. reicht ja ein Histogramm, Boxplots etc.) auch aus.

- Auch wenn man eine Programmiersprache für die Datenanalyse benutzt, können nicht alle Analysen bzw. ihre Skripte standarisiert geschrieben werden, weil vieles darin besteht, unterschiedliche mögliche Wege auszuprobieren. Am Ende entsteht dadurch kein Skript, sondern eher eine “Geschichte” der Auswertungswege, die man getestet hat. Bei anderen Aufgaben ist es leichter, die Vorgänge in Skripten zu automatisieren
- Würde ich mir persönlich wünsche, die Frage ist nur wann? Vielleicht hier wirklich auch mal einen gezielten Workshop oder auch eine “Übung” über mehrer Wochen geblockt
- hängt von der Aufgabe uns Art von Daten ab

Daten visualisieren, IST-Zustand

- Ich kann mich hier schlecht einordnen. Eine gute einfache Grafik kann wichtiger sein als eine interaktive Spielerei.
- Im Moment werden Visualisierungen nur mit excel erstellt, bzw. erstellen Kollegen diese mit R.
- Visualisierung in Excel, da hohe Flexibilität und schnelle grafische Änderungen gefragt sind. Keine Darstellung von Unsicherheiten möglich, oder nur umständlich realisierbar. Zusammenfassung einzelner Datensätze zu Blöcken sehr individuell, daher Visualisierung nicht nach Standard möglich.
- bei fortlaufenden Daten geht es nicht anders. Für Präsentation gehe ich zu Punkt 1 über
- mit R

Daten visualisieren, SOLL-Zustand

- Ich würde gerne nachvollziehen können, welche Grafikdatei von wem/welchem Skript erzeugt wurde. Gut wäre, Ergebnisgrafiken je Projekt an einem einheitlichen Ort abzulegen, da sie am ehesten für Präsentationen benötigt werden.
- um nach Ebene 4 zu gelangen, müsste ich dann immer einen Programmierer im Projekt haben, da ich bezweifle dass ein Skript alle Anforderungen für jedes Projekt erfüllt, auch ist vielen Leuten nicht deutlich welche Möglichkeiten der Visualisierung existieren
- Auf Grund der Datenmenge wird ein anderes Tool als excel notwendig.
- Darstellung von Unsicherheiten sollte verbessert werden, aber die Rohdaten lassen das nur schwierig zu. Ausgabe von einzelnen Modellzuständen mit unfangreichen Daten, daher keine automatische Modellierung von Sensitivitäten oder Unsicherheiten möglich (oder noch nicht ausprobiert!).
- Da ich mich persönlich nicht in R einarbeiten werden, sind alle Darstellungen in R für mich persönlich nicht veränderbar. Oft will ich einfach vor einem Vortrag auf einer Tagung eine Abbildung ändern z.B. Achsenbeschriftung/Größe etc. und es ist sehr umständlich erst die MA darum zu bitten. (vorausgesetzt ich finde die Exceldatei mit den aktuellen Labordaten und der Grafik, die ich ändern möchte.) Die Datenvisualisierung sollte so wenig Aufwand wie nötig sein, für Labordaten oder z.B. Output von Umberto reichen Execl/Origin völlig aus. R kann helfen die Rohdaten aus mehreren Datei (z.B. Labordaten BWB) für die Auswertung zusammenzuführen.
- Ich bin der Meinung, Visualisuerungen, die der Auswertung dienen, müssen nicht immer interaktiv sein
- Ich würde gerne über die "Excel" hinauskommen, aber dafür auch Unterstützung um es auch zu "lernen"
- es ist eigentlich egal von wem die Tools programmiert werden

Daten kommunizieren, IST-Zustand

- Ich versuche wenn es das Projekt zulässt so transparent wie möglich zu arbeiten wie beispielsweise im Projekt DEMOWARE. D.h. das R Paket auf Github/Zenodo veröffentlicht, zusätzlich die der Risikoberechnung zugrunde liegenden R-Skripte auf Zenodo. Zenodo bietet ist zudem in das OpenAIRE System der EU eingegliedert und erlaubt die Verlinkung mit dem "Grant"
- Die Ergebnisse werden als Präsentation oder Bericht kommuniziert.
- Nur statische Visualisierung, einfach, erprobt und schnell verständlich (Säulen- oder Balkendiagramme in Excel, ggf. Unsicherheitsbalken).
- Ich kommuniziere soweit es nötig und sinnvoll ist.

Daten kommunizieren, SOLL-Zustand

- Ich würde es wünschenswert finden wenn die Geschäftsführung des KWB ein Statement zur transparenten Ergebniskommunikation abgeben würde (wo immer das möglich ist). Persönlich denke ich, dass es das Projekt FAKIN gerade auch deswegen gibt, da wir eigeninitiativ in öffentlichen Repositorien (Github, Zenodo) veröffentlicht haben. Nun wäre die Frage ob so ein Vorgehen auch in anderen Projekten möglicherweise positive Effekte hätte?
- Ich denke, es wäre gut, auch Projektzwischenergebnisse auf der Homepage zu veröffentlichen. Eine professionelle Aufbereitung für die Öffentlichkeit (Vergabe an Profis?) wäre wünschenswert. Eigene Daten sollten wir am Ende eines Projekts in öffentlichen Repositorien zugänglich machen.
- öffentliche Repositorien stellen die Frage nach Urheberrecht
- Ggf. könnten neuen Methoden der Darstellung und Kommunikation die Kommunikation verbessern. Rohdaten und Berechnungen sollten nicht öffentlich zugänglich sein, da Daten immer vertraulich geliefert werden und erst in Berichtsform nach Zustimmung der Partner veröffentlicht werden.
- Die Rohdaten z.B. von BWB sind ggf. nicht für Öffentlichkeit bestimmt.

Datenkultur, IST-Zustand

- Am KWB ist bereits viel Know-How im Umgang mit Daten vorhanden. Dieses ist auf viele Mitarbeiter verteilt.
- Kein System zur Datenverarbeitung vorhanden, aber viele Mitarbeiter die sich eigene Methoden entwickelt haben. Sehr vielfältige Datenformate und Anforderungen, daher individuelle Lösungen für verschiedene Datentypen notwendig.

Datenkultur, SOLL-Zustand

- Gerade die IT-Ressourcen bereiten oftmals bei der Arbeit mit großen Datenmengen Probleme. Wohin mit den vielen Rohdaten, wenn der Speicherplatz auf "POSEIDON/PROJEKTE" sowieso schon fast immer voll ist. Zudem stellen die immer größeren Datenmengen (mehrere Millionen Datenwerte) insbesondere häufig die Praktikanten Computer mit zu wenig Arbeitsspeicher (< 8 GB) oftmals vor Probleme. Positiv ist, dass die Geschäftsführung die Datenkompetenz-Weiterbildung aktiv durch die Finanzierung eines DataCamp Accounts zum erlernen von R, Python und SQL fördert.
- Es gilt, das am KWB vorhandenen Wissen über Datenmanagement zu nutzen, indem es zusammengetragen, vereinheitlicht und regelmäßig kommuniziert wird. KWB-interne Standards (z.B. Namenskonventionen) sollten definiert werden. In Projektanträgen sollte "Datenmanagment" konsequent mit ausreichend Personal- und Sachmitteln (z.B. für Server, Speicher, Rechenkapazität) eingeplant werden.
- Das QM sollte einen Rahmen vorgeben, ohne zu sehr einzuschränken
- Datenmanagement könnte deutlich verbessert werden, gerade Aufbereitung/Darstellung und auch Archivierung. Datenanalyse sollte individuell angepasst werden und je nach Anforderung gestaltet werden (z.B. Excel, Origin, R, ...).
- Meiner Ansicht nach ist Datennutzung ist kein Selbstzweck
- Daten und deren Auswertung sind der Kern wissenschaftlicher Arbeit und der Umgang mit ihnen soll Teil einer Verbesserungkultur sein.
- das wird einmal zu Beginn druchgelesen und unterschrieben, aber es wird nicht mehr darauf hingewiesen. Ich würde mir hier auch von den Verantwortlichen regelmäßig Information bzw. auch Möglichkeiten zur "Weiterbildung" wünschen (kann ja vielleicht auch als "interne Fortbildung" gesehen werden, um so Kosten zu sparen)
- die Datenkompetenz muss innerhalb der Projekte vorhanden sein. Wenn eine Weiterbildung dafür notwendig ist, ist sie bei demJahresgespräch zu identifizieren, wie alle andere Weiterbildungen.

Urheberrecht und Datenschutz, IST-Zustand

- Wir haben bereits IT-Richtlinien.
- Meine Rohdaten werden grundsätzlich nicht an extern weitergegeben oder veröffentlicht. Aufbereitete Daten werden nach Freigabe durch den Partner in Berichtsform veröffentlicht.
- Ist ein heikles Thema, da man mit den verschiedensten Arten Von Daten zu tun hat (Fotos, Graphiken für Präsentationen, Daten dritter,....)

Urheberrecht und Datenschutz, SOLL-Zustand

- Sollten wir die existierende IT-Richtlinien ergänzen? Sobald wir Daten von Externen erhalten, sollten wir explizit (standardisiert) abfragen, welche Nutzungsbeschränkungen es gibt und diese Information standardisiert protokollieren und ablegen.
- Haben Daten einen Urheber? Wenn ich beispielsweise den pH Wert in der Havel im Auftrag der BWB messe "gehört" diese Messung dann dem KWB, den BWB?
- Grundsätze für Datenschutz sollten allen bekannt sein und in allen Projekten gleich angewandt.

- Wär ein Wunsch, um sich auch rechtlich abzusichern, aber es ist ein bürokratischer Aufwand wo man sich die Frage “Aufwand/Nutzen” besser (“Aufwand/Schaden”) stellen muss.