Kommentare zur Version vom 1. April 2019

a) Markus Sch.

I) Aus meiner Sicht unbedingt notwendig sind Änderungen bei:
a) In der Tabelle auf dem Deckblatt steht „on behalf of Commitee“
Es war der explizite Wunsch von KET, dass dies so nicht erscheint.
Der Ausdruck spiegelt nach Meinung mehrerer Leute auch nicht den Prozessverlauf wieder.
→ Wird im Treffen der Komiteevorsitzenden mit Editorenteam am 10.4. diskutiert.
b) Abschnitt 5 Zeile 706 bis 708:
„extensive survey … in-depth analysis …“
Das hat meiner Meinung nicht stattgefunden und sollte so definitiv auch nicht erscheinen.
Ich schlage vor, dass hier ein andere Formulierung gewählt wird.

II) Es wäre wünschenswert auch folgende Passagen noch klarer zu gestalten:
a) bei 5.4. steht explizit long-term position" aber z.B. bei 5.2 nicht.
Was genau ist mit „long-term position“ gemeint?
Warum wird hier ein Unterschied in der Bezeichnung in 5.4. und an anderer Stelle z.B. 5.2 gemacht? Ist die Unterscheidung beabsichtigt und falls ja, wie ist der Unterschied begründet?
b) an mehreren Stellen steht "ErUM-Communities including physics, medicine, biology, chemistry“ (z.B. Zeilen 1041 und 1159,). Das ist meiner Meinung nach nicht korrekt. ErUM umfasst im wesentlichen die 8 Komitees und nicht auch die anderen Forscher*innegemeinden. Ich fände eine andere Formulierung wünschenswert.
→ Thomas Kuhr hat geantwortet und ist für MS nicht wichtig.

Keine Kommentare zu Ausdrucksweise und Typos, Formatierung, da noch von Muttersprachler geprüft wird.

b) Stefan Kluth (zweite E-Mail mit konkreten Vorschlägen)

Generell, ueberall wo "federated infratstructures" steht, nur "infrastructures" schreiben. Wir haben keine Definition des Begriffs "federated infrastructure" und so besteht die Gefahr das jede community das anders versteht. Und wir sollten nicht die Conclusion (das wir federated infrastructers (im HEP Sinn) brauchen) an den Anfang stellen, bevor wir die Fakten gesammelt und ausgewertet haben. MMn kommt am Ende raus, dass das Arbeiten mit Exabytes von Daten mit Big Data Analytics training Anforderungen nur von grossen Infrastrukturen geleistet werden kann. Ob die neu oder als "Erweiterung" vorhandener Anlagen entstehen ist "nicht unser Problem hier" sondern soll mit unserem ErUM Input im Rahmen von z.B. NFDI beantwortet werden.

l 52 ff: New funding schemes are necessary to enable this development, thus allowing to exploit the full potential of the data provided by the next generation facilities. Such changes must be well coordinated with other programs concentrating on establishing novel infrastructures such as the National Research Data Infrastructure (NFDI) and the European Open Science Cloud (EOSC). Our work on infrastructures for future ErUM research in a evolving world of Big Data and Big Data Analytics can thus be seen and used as input by the ErUM community for the development of the NFDI.

l 76 ff ~~Federated~~-Infrastructures: provisioning of ~~federated~~ infrastructures "beyond Moore"

l 370 ff Ein extra bullet:
o consideration of the special demands of training of Big Data Analytics tools with the expected data volumes in the Exabyte range

l 403 ff bzw l 405 ff nur ein bullet (die beiden vorhandenen koennen auch widerspruechlich gelesen werden)

o Creation of a federation of computer centres serving as a "ErUM Science Cloud" with integration into national concepts such as the "National Research Data Infrastructure" as well as EU-funded endeavors such as EOSC.

l 419 ff The infrastructure needs to meet the future challenges in terms of storage, (German share of data volumes of exabyte per year), network (high-bandwidth network with 100 gigabit per second short-term and terabit per second long-term) connecting all ErUM centres, computing power for data processing, simulation and Big Data Analytics tools training, and services.

l 727 ff o ~~Federated~~ Infrastructures: Owing to the rapid increase in data volumes and the rise of Big Data Analytics tools in the coming years, the demand for compute power, storage space and networks will increase dramatically. We see an immediate need for action in the well-planned and significant expansion of the infrastructure.

l 786 am Ende des Absatzes dazu: In addition, the training of Big Data Analytics tools on massive amounts of data, which is a rapidly evolving field of research, will very likely need very large and specialised compute resources.

Erste E-Mail:
ein paar kleine Edits:
l 233: " ... model in a concerted action of university, HGF and Max-Planck data centers." (We have Tier-2/2 at MPCDF, and use HPC there for ATLAS simulation.)
l 258: "Scientists with in depth ... " note plural
l 263: "scientists" (plural)

Ich habe auch einige Beobachtungen zu dem Teil ueber Infrastrukturen. Mir ist der intendierte Plan des Write-ups auch erst nach einigen Iterationen klar geworden. Ich weiss das jetzt nicht mehr gross Umgeschrieben werden soll, aber einiger Punkte sind nur kleine Aenderungen, oder sie dienen als Hintergrund fuer weitere Iterationen der Konzepte.

l 342-350: Hier beschreiben wir die Probleme des WLCG Computing Models: "Such a heterogeneous environment poses significant challenges for the design of future computing models, software development and workflow and data management.". Dann sagen wir (implizit), das wir das weiter machen wollen, trotz der Probleme.

Wie wollen wir die moegliche Kritik beantworten, das wir das "WLCG-artige" Computing Model nicht hinter uns lassen wollen, obwohl es die kuenftigen Bedarfe nicht effizient erfuellen kann (wenn es jetzt schon Probleme macht). Warum planen wir nicht ein veraendertes Computing Model (so wie z.T. es auch im HL-LHC Umkreis gemacht wird), mit viel weniger und dafuer grossen Standorten, die dann aufgrund von economies of scale signifikant effizienter betrieben werden koennen.
Unsere Aufgabe waere es dann, die Workflows und Software unserer Experimente entsprechend anzupassen.

Im naechsten Absatz wird ueber notwendige Entwicklungen fuer science clouds gesprochen. Nur wirkt das fuer mich widerspruechlich wg des Festhaltens am WLCG-artigen computing model. Natuerlich muessen wir diese Entwicklungen machen, die sind sogar absolut notwendig um die neuen Infrastrukturen nutzen zu koennen. Erst wenn wir klar machen, das wir auf ein grundlegend neues und im Vergleich (mit WLCG) besseres Computing Model umsteigen wollen (muessen) wird die Entwicklung entsprechender Workflows und Software zwingend notwendig.

l 368 ff: requirements for infrastructures
Nachdem die ueberragende Wichtigkeit von "big data analytics" und den dafuer benoetigten compute resourcen (grosse GPU faehige cluster um z.B. deep neural networks mit sehr grossen Datenmengen zu trainieren) diskutiert wurden, schreiben wir diese Anforderung hier nicht in die Liste. Ich denke es ist klar das grosse spezialisierte (d.h. GPU faehige) HPC Systeme gebraucht werden, um das Training

von analytics tools zu ermoeglichen. Um damit (big data analytics) wirklich Erfolg zu haben, wird der unregelmaesige und unsichere Zugriff auf z.B. Gauss-Allinazsysteme oder ein paar Server mit GPUs an den Standorten nicht ausreichen.

l 327 NFDI (und EOSC)
Wir erwaehnen diese Projekte hier um mitzuteilen das wir Bescheid wissen. Ich meine aber wir sollten auch klar Stellung beziehen, wie die Zusammenarbeit aussehen soll. Das koennte die klare Trennung von Infrastrukturaufgaben in NFDI (und EOSC) und allen anderen Aufgaben (Software, Workflows, Ausbildung, wissenschaftliche Arbeit mit den Daten), die hier besprochen und dann auch beantragt werden. Dazu muesste man sagen, das "wir" (ERuM) bei der Gestaltung der NFDI fuehrend mitwirken wollen und muessen, und dafuer die Ergebnisse der WG 1 einbringen. So wuerde zum Ausdruck kommen, das wir eine klare Vorstellung von der Verzahnung der verschiedenen Intiativen haben.

c) Markus Elsing

content:
---
l.196: term “Smart Data” is not introduced and the context here therefore stays vague. Either drop this or introduce this term with a definition of what this should mean in the context here. → Nein

l.232: MPI as well provides computing infrastructure, should we add this ? →Ja

l.608: I do have some issues with calling "but also the analysis tools and used libraries” inside "meta data”. I would have the tendency to drop this half-sentence. Analysis tools and libraries are provided on most systems by other means than (meta) data. Systems like CVMFS and alike are used for this, and so are container solutions. I guess this is a questions of where to put the boundary… → Nein

l.616: the first bullet gives the impression that FAIR “only” means a DOI. A DOI is part of making the data FAIR, but that is obviously not sufficient. I would use the DOI as an example of what is required and write it like this.

l.619: I would reorder the bullets, putting the third one as second in the list.

l.808ff: Here we only mention HPC as a system providing computing resources. It provides very powerful resources of a certain type, but we should mention this at the same level as HTC and classical GRID/cloud type of resources, not to give the impression we would favour a concentration on HPC investments.

Section 5 reads now much better to justify the resource needs. But 5.2 (l.899 ff) does not mention that workflow and data management services are developed in an EU and international context, we had e.g. mentioned EU ESCAPE project (DOMA in particular) in the text above section 5. We can/should benefit from the EU/international collaboration across domains (a la ESCAPE) and aim at even take leading roles where appropriate.

Even more so for section 5.3 (part about challenges and responsibilities), which is maybe not the strongest in terms of justification. I would prefer to talk as well here about the development, deployment and operation of services for data management and workflow management. Like in section 5.2, the EU and international context for such developments are missing, ESACPE/DOMA and e.g. Rucio as a system used by more and more of our communities.

l.966: Oh, this I really don’t like “committees to design the data” - this never ever worked for any cross community projects I am aware of. Data structures need to be transparent and appropriate for the domain problem. Experiments will deal with this, not an external committee that is bound to lack

experiment specific expertise. For cross-experiment data sharing one needs to make data structures public and usable, not force experiments to use structures that don't match their needs.

l.1042 and l.1159: I would also drop "including physics, chemistry, biology, medicine, earth science" as others commented already, this is not precise/misleading and not needed here.

l.1165: "brain drain" is not what I would write here, it is a goal that we educate data scientists for industry. But the balance is not there, we risk to loose the experts faster than we can educate new ones. I would formulate a bit more careful.

l.1320: I thought we do not include Fair TIer-0 and SKA resources in the numbers here, but then it should be said ?

minor text issues:

Formatting issue with line breaks:
l.71, 100, 347, 572, 914, 934, 1079
l.176 "… dramatically, exceeding …" (comma)
footnote on page 1 is same as in executive summary, drop.
l.313 "… centres, over"
l.315 "… clouds, to"
l.339 drop "(Moores law)" - was already introduced before
l.447: "… data rate, but also"
l.526 "… computer science, while"
l.527/8: "collaborative efforts" -> "them" (duplication in sentence)
l.535/536 sentence reads odd to me, can you check it ?
l.549: drop "suggested"
l.577: drop "(data mining)" as term is not introduced before and does not add to the context here
l.599/600: "data types" is 2 times in the same sentence, please fix wording
l.887: The sentence is complicated, I would write "processing; on the other hand
l.984: "measure" reads strange to me, not sure this is proper english