

Dear Editors and Reviewers of the Journal of Information and Data Management,

We would like to thank you for the helpful comments and suggestions concerning our manuscript entitled "Multiscaling a Graph-based Dataspace". We believe our article has improved significantly thanks to such feedback. We decided to do a major modification on the structure of the paper.

In general, we added a new section to accommodate formalization aspects including a multiscale graph model and the transformation process, providing details about the approach. We also further describe the implementation aspects and provide additional details in experimenting our work in the Organism-Centric Analysis. This allowed us to better discuss the obtained findings. In addition, we improved the readability and clarity as well as typographical and language issues addressing the minor essential revisions. More specifically, we described sentences more directly to avoid misunderstandings and revised the English language. The following changes and responses refer directly, point-by-point, to the issues raised by the reviewers.

Best Regards,
Authors

1. Reviewer #1

1.1. Reviewer #1 - Comment A

C. "I missed a formalization section"

A. We added a section with a more formal definition regarding the underlying model and the transformation processes. It is entitled "Multiscale Graph Model".

1.2. Reviewer #1 - Comment B and C

C. "I missed: a discussion about the decisions made: why to use a graph model; how to solve the semantic interoperability; A discussion on the models used to represent the graphs at each layer;"

A. A formal graph model representation was added to clarify the data elements in each layer and how the transformations from one scale to other occurs. We improved the justification regarding our design choices. We explore graphs due to the flexibility to represent different structures along the scales (this was highlighted in Section 3).

The semantic interoperability was discussed in more details in the Conceptual Scale (Section 3) and in the transformation of the Description to the Conceptual layer (Section 5.6). We discuss aspects like entity and property representation and explicit semantics.

We present a discussion of the models in each layer in two sections: Section 3 (presentation of the data architecture in general terms) and in Section 5 (presentation of the data architecture in the biology scenario). We improved the presentation and better interconnected both sections.

1.3. Reviewer #1 - Comment D

C. "why does a description scale is needed? Could it be part of the conceptual one, what are the benefits of the chosen architecture?"

A. The definitions of the scales were properly inspired by empirical previous work that indicated their need. To clarify this question, we highlighted in the text that the Description scale focus on providing a single logical model for further processes. We maintain the Description Scale differentiated from the Conceptual Scale since algorithms to discover and extract semantics from content are evolving and can be domain specific. Therefore, in a layered architecture all these algorithms will depart from the same description scale, focusing in their specific role, producing different conceptual scales. The benefits of our architecture are shown in the experimented biological scenario presented in section 5, where we improved the distinction of the two scales and their importance.

1.4. Reviewer #1 - Comment E

C. "Figure 3 shows a visualization of the description scale. How does it get produced? Is it part of the contribution of this paper?"

A. Figure 3 (now Figure 11) is a result of a software produced by us which gets data from the description scale to visually show some preliminary results to the biologists. It was mentioned for the first time in the short paper of the SBBD 2015 -- this paper is an extension of the SBBD paper.

1.5. Reviewer #1 - Comment F

C. "i found the approach highly ad-hoc, although the authors claim to be systemic. There is not a single algorithm that systemically explains how things get done. Please, show us the algorithms for scale translations;"

A. We now provide the formalization and algorithms for scale transformation to support our claim of systematization. However, systematization also refers to the idea of distinguishing integration steps and organizing them in progressive steps. This organization, we call now Primary Data Architecture, gives to the user/developer a reference set of incremental steps to follow.

1.6. Reviewer #1 - Comment H

C. "there is no validation in the paper. how can we trust that this approach leads to a "any criteria" a quality solution."

A. We reorganized our implementation description in Section 5 to better illustrate how our model is reflected in concrete implementations. Even though we still do not have a single unified system that implements the complete pipeline, every scale and transformation in the system was implemented independently, as discussed in Section 5.1. We are now integrating all the independent implementations in an unified system, see the progress at: <http://linkedscales.lis.ic.unicamp.br>. This system might provide further evidences of the outcome quality from our proposal.

2. Reviewer #2

2.1. Reviewer #2 - Comment A

C. "Seção 4. Analisar se os dois primeiros parágrafos dessa seção devem ser colocados em outra seção, talvez a seção 3. Daí, a seção 4 seria voltada apenas à proposta do LinkedScales, e não mais à motivação relacionada. Acredito que fique melhor dessa forma."

A. Concordamos com a sugestão e fizemos as alterações. Mantemos uma parte do parágrafo apenas com o intuito de fazer uma ponte entre a seção anterior e seção atual, mantendo a redação um pouco mais fluída. O paragrafo 2 foi dissolvido na seção 4.

2.2. Reviewer #2 - Comment B

C. "Seção 4. Apesar de repetitivo no texto, porque já está destacado no abstract e na introdução, acrescentar frases iniciais antes do terceiro parágrafo dessa seção para dizer que neste artigo é proposto LinkedScales, Ou seja, falta uma "apresentação" do LinkedScales na seção, antes dele ser descrito."

A. Concordamos com a sugestão e fizemos alterações na apresentação da seção.

2.3. Reviewer #2 - Comment C

C. "Seção 4. Analisar se o comentário referente a "... based on previous experiences on data integration..." deveria ser mudado para a seção de trabalhos correlatos. Nesse ponto do texto, tudo indica que o comentário deva ficar em trabalhos correlatos e, em trabalhos correlatos, colocar o que tem diferente para o trabalho atual."

A. Alteramos o termo "based on" por "inspired by". A intenção aqui era dizer que nos baseamos em experiências anteriores, e não que este trabalho seja uma extensão ou subproduto dos trabalhos desenvolvidos anteriormente.

2.4. Reviewer #2 - Comment D

C. "Seção 4. O último parágrafo da seção deve ser posicionado antes de se explicar a escala conceitual, porque ele se refere à escala descritiva. Ainda sobre esse parágrafo, os autores devem explicar o exemplo um pouco melhor, para diminuir a dificuldade do leitor em entender o que está sendo explicado."

A. O artigo sofreu uma reestruturação e melhorias gerais para clarificar os parágrafos. Acreditamos que nesta nova versão este problema foi resolvido. Conforme solicitado, o parágrafo e a respectiva figura foram incorporado à descrição da escala descritiva.

2.5. Reviewer #2 - Comment E

C. “As explicações no texto estão em termos de cores, mas eu imprimi o artigo usando uma impressora branco e preto e daí não acompanhei os exemplos em cores. Alterar essas referências a cores no texto.”

A. Modificações nas figuras foram elaboradas para melhorar esses aspectos de leitura do artigo. Entretanto, encontramos dificuldades em não usar referências de cores em algumas figuras complexas que envolvem sobreposição de indicadores.

2.6. Reviewer #2 - Comment F

C. “Melhorar a legibilidade das Figuras 6 e 7.”

A. Modificamos a figura para uma outra com uma resolução melhor.

2.7. Reviewer #2 - Comment G

C. “Uma forma de melhorar a legibilidade da explicação da Figura 8 é colocar (a) (b) (c) e (d) em cada uma das escalas representadas nessa figura. Daí, no texto, ao se falar da escala física, por exemplo, usa-se Figura 8(a). Isso facilita o entendimento do texto. A mesma observação é válida para as Figuras 10 e 11.”

A. Modificamos a figura e texto de acordo com as sugestões.

2.8. Reviewer #2 - Comment H

C. “Seção 5.5. Faltou um último parágrafo nessa seção apresentando alguma conclusão ou comentário a respeito da implementação. Quais os resultados obtidos com a implementação realizada? É possível determinar como a implementação contribuiu para se analisar alguma característica da arquitetura proposta?”

A. Nós adicionamos novos detalhes sobre a implementação e integramos sua apresentação com o detalhamento do experimento prático. Isso nos permitiu melhor discutir os experimentos e resultados com base no protótipo alcançado.

2.9. Reviewer #2 - Comment I

C. “Pergunta. No início do texto, descreve-se várias vezes que os esquemas das fontes podem sofrer alterações, e que a abordagem baseada em dataspace é mais indicada porque é mais flexível. Como a arquitetura proposta é usada frente a alterações nos esquemas das fontes? Isso não está claro no texto.”

A. A estrutura/grafos ortogonais que existe entre as escalas permite rastrear as operações

realizadas a partir da representação da fonte na escala física. Apesar de não atacarmos este problema em específico, acreditamos que essa estrutura permite que as transformações sejam refeitas ou refatoradas (dependendo da modificação que aconteça). A nova versão visou deixar este aspecto mais claro.

2.10. Reviewer #1 - Comment J

C. “Os comentários a seguir consistem em comentários mais específicos de escrita:

1. Abstract. Quando se diz “... this work discusses the conception ...”

Usar presents ou introduces ao invés de discusses.

2. Seção 1. Colocar referências no final da primeira frase do segundo parágrafo. Ou seja, colocar referências para o termo Dataspaces.

3. Seção 2. A sigla GS foi introduzida no primeiro parágrafo dessa seção e usada somente uma única vez nesse primeiro parágrafo. Portanto, não há a necessidade de se manter o uso dessa sigla.

4. Seção 3. Verificar se a escrita da última palavra da terceira linha do segundo parágrafo está correta.”

A. Essas modificações específicas foram feitas na nova versão do artigo.

3. Reviewer #3

3.1. Reviewer #3 - Comment A

C. “In Section 2, the authors claim that the predominant approach for data integration is based on the concept of a global schema. First of all, the idea of a predominant data integration approach is not correct. There are several approaches, such as federated databases, multi-databases, schema integration and data warehouses. Moreover, federated databases and multi-databases do not require the concept of a global schema. Such approaches, together with the mediator strategy, should be discussed in Section 2.”

A. In fact, the text is inclined to define global schema as a “classical”/predominant approach for data integration. Our idea, however, was to make a discussion about the upfront strategies vs. the incremental ones, being the definition of a global schema a well adopted strategy. We made changes aligned with your comment. Furthermore, we avoided the “classical” vs. modern. Rather, we emphasized that there are several upfront costly strategies, including global schemas, federated databases, multi-databases, schema integration and data warehouses.

3.2. Reviewer #3 - Comment B

C. “It is not clear how a source database is mapped to a graph database. For example, it should be explained how the data of an XLS file can be mapped to a graph. Besides, the authors do not mention if this process is carried out without human intervention and if there is formal process for such a task.”

A. This mapping process is carried by a specialized software that we developed, whose name is 2Graph. It works automatically by mapping the sources to the graph. We did not give details because it is not the focus of this paper. Nevertheless, we clarified the role of the 2Graph in the paper, and we also mentioned its automatic mapping, pointing to a paper in which we describe in details the previous version of 2Graph (named DDEx).

3.3. Reviewer #3 - Comment C

C. “In my opinion, the Description Scale presents the same features of a global.”

A. We addressed this concern in Issue 1.3. Please refer to the section **Reviewer #1 - Comment D** for details.

3.4. Reviewer #3 - Comment D

C. “The paper does not present any experimental evaluation of the proposed approach. For instance, is LinkScales scalable regarding the size of data sources and the amount of data sources?”

A. We addressed the concern about experimental evaluation in Issue 1.6. Please refer to the section **Reviewer #1 - Comment H** for details.

Regarding the scalability of size and amount of data sources, additional and specific evaluation aspects like size and scalability were point out as future work.