Lexicographic Data Seal of Compliance

Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Irena Marinski

▶ **To cite this version:**

Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Irena Marinski. Lexicographic Data Seal of Compliance. [Research Report] ELEXIS; DARIAH. 2021. hal-03344267

HAL Id: hal-03344267
https://hal.science/hal-03344267

Submitted on 14 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

elexis european lexicographic infrastructure

Lexicographic Data Seal of Compliance

Author(s): Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Irena Marinski

Date: 30 July 2021

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017
Grant Agreement No. 731015
ELEXIS - European Lexicographic Infrastructure

Lexicographic Data Seal of Compliance

Deliverable Number: D6.4

Dissemination Level: Public

Delivery Date: July 30, 2021

Version: 1.0

Author(s): Toma Tasovac, Laurent Romary, Erzsébet Tóth-Czifra, Irena Marinski

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

elexis european lexicographic infrastructure

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Gran Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by
V1.0 30/07/2021	Submission-ready	Toma Tasovac

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Table of Contents

Acknowledgements

The authors of this proposal would like to express their thanks to members of the DARIAH Working Group “Lexical Resources”, the ELEXIS Standards Committee and the ELEXIS Integration and Sustainability Committee for their feedback on earlier versions of this document.

Scope

The Lexicographic Data Seal of Compliance (LexSeal) is a proposal for a community-based certificate of compliance with best scholarly practices to be awarded to individual *lexicographic datasets* in recognition of their creators' self-assessed and well-documented adherence to the principles of *trustworthiness*, *interoperability*, *stewardship, citability, reciprocity and openness*.[1]

These six dimensions of LexSeal taken together describe the *infrastructural fitness-for-purpose* of a given lexicographic dataset, i.e. the degree to which the dataset aligns with best practices and technical standards within a larger network of machine- and human-readable lexicographic data.

What is a lexicographic dataset?
A lexicographic dataset is a machine-readable lexicographic resource. Machine readable here means that the text of the dataset can be reliably processed, extracted and manipulated as text. Scanned images of lexicographic resources lie outside the scope of LexSeal.

What is a lexicographic resource?
A lexicographic resource is a type of information resource which lists and/or describes some or all the words in one or more languages, regardless of its specific purpose (documenting a historical, contemporary or an endangered language; helping learners learn a new language, enabling part-of-speech tagging and other kinds of NLP annotations etc.)

Who awards the certificate?
The certificate shall be awarded by a cross-institutional governing body, preferably under the auspices of ESFRI Landmark Research Infrastructures DARIAH and CLARIN, as well as the future ELEXIS Association.

Why certify in the first place?
The creation and curation of digital datasets, in general, and lexicographic datasets, in particular, requires time, effort and professional skills. Exploring existing datasets is an integral part of research discovery, while sharing one’s own lexicographic datasets can lead to better returns on one’s own professional and institutional investment by establishing

[1] These principles have been put forward and validated by communities around the Heritage Data Reuse Charter https://datacharter.hypotheses.org/. See also Laurent Romary and Erzsébet Tóth-Czifra. 2019. Open Access guidelines for the arts and humanities: Recommendations by the DARIAH European research infrastructure consortium. ⟨halshs-02106332⟩

connections and encouraging collaboration with other dataset creators. Certification as a quality assessment mechanism has a double purpose: it facilitates successful, efficient and productive data-sharing practices, and functions as a reward system for dataset creators.[2]

LexSeal is a trust-based system aimed at three groups of stakeholders: creators and dataset providers, funders and dataset users. Each group can benefit from certification in different ways:

- **creators and dataset providers** can show to both their users and their funders that an independent authority has validated, evaluated and endorsed the overall infrastructural fitness-for-purpose of a given dataset; the visibility and citability of certified datasets could contribute to their creators' scholarly and professional prestige, whereas the certification process itself could become a mechanism for intellectual exchange on best practices in the domains of lexicography and lexicographic infrastructures;
- **funders** can use the certification mechanism as a way of ensuring that the funded lexicographic datasets meet a set of community-based best practices and criteria, and, by extension, that their investment was well spent, resulting in the creation of interoperable resources that are well-documented; and, finally,
- **dataset users** can more easily assess the extent to which a given dataset meets their particular scholarly or professional needs, the degree to which it adheres to best practices and the conditions under which it can be reused; in addition, they can find out who to get in touch with in case they have questions or suggestions.

Why certify in the long run?
The proliferation of community-certified lexicographic datasets can have a positive effect on the solidification and expansion of a sustainable data-sharing ecosystem by giving additional visibility to existing resources and by encouraging best practices in the creation of new resources.

What LexSeal is not?

- LexSeal is *not* a repository certification system. Those already exist (see Background below). We focus on individual datasets because we want to facilitate access to and reuse of specific lexicographic datasets while recognizing their inherent richness and diversity. Stable hosting of lexicographic datasets is only one of the parameters to be evaluated as part of the LexSeal certification process.
- LexSeal is *not* a data representation standard. It is a certification mechanism that builds upon the existing standardization activities (TEI, ISO, W3C etc.) and evaluates them in the context of overall accessibility, reusability and sustainability of lexicographic datasets.
- LexSeal passes *no* judgement on the methodological soundness of a given lexicographic resource, the quality of its linguistic content or the professional expertise of its creators. As such, it is *no* replacement for the scholarly assessment of lexicographic resources which one encounters in the scholarly literature (journal articles, book reviews etc.), but rather complementary to it.

[2] See Edmond, Jennifer, & Tóth-Czifra, Erzsébet. 2018. Open Data for Humanists, A Pragmatic Guide. Zenodo. http://doi.org/10.5281/zenodo.2657248; and Eve, Martin Paul. 2020. 'Violins in the Subway: Scarcity Correlations, Evaluative Cultures, and Disciplinary Authority in the Digital Humanities'. In *Digital Technology and the Practices of Humanities Research*, edited by Jennifer Edmond. Cambridge: Open Book Publishers.

Background

LexSeal has been long in the making. The discussions around a certification mechanism for lexicographic datasets were initiated by Toma Tasovac and Laurent Romary in the context of the DARIAH Working Group “Lexical Resources” against the background of a number of existing initiatives and trends in the scholarly communities and research infrastructures ranging from standards and FAIR principles to certified data repositories and the Heritage Data Reuse Charter.

Standards
Various *standards* and *data formats* are increasingly affecting the workflows of researchers and end users, and not only those directly involved in the development and application of computational methods. Professional and amateur lexicographers do not have to be computational linguists or NLP experts to use generic dictionary writing systems, such as Lexonomy,[3] or dictionary viewers, such as LEX2[4], but they do need to be aware of what structured data is, why it matters, how to create it and how to use it. They also need reliable information *about* lexicographic datasets so that they can more easily assess their usefulness in a given context.

FAIR Principles
The FAIR principles for scientific data management and stewardship[5] provide guidelines for improving machine-actionability (i.e. the capacity of computational systems to process and interact with data with no or minimal human intervention) in terms of their *findability* (via machine-readable metadata), *accessibility* (via the use of standardized communication protocols), interoperability (by being open to integration with other data or workflows for analysis, storage or processing); and *reusability* (via well-described metadata attributes and clear data usage licensing). While the FAIR principles have received wide political support[6] and are well-established in science, technology and innovation domains,[7] there is still significant room for improvement:

---

[3] Měchura, M. B. 2017. ‘Introducing Lexonomy: an open-source dictionary writing and publishing system’ in *Electronic Lexicography in the 21st Century: Lexicography from Scratch*. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.

[4] LEX2 is currently being developed as part of the ELEXIS project.

[5] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

[6] See, for instance, the G20 Leaders’ Communique from the Hangzhou Summit https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967; also: European Commission, Directorate-General for Research & Innovation, H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 (26 July 2016), http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf ; Australian FAIR Access Working Group, Policy Statement on FAIR Access to Australia’s Research Outputs, https://www.fair-access.net.au/fair-statement; Erzsébet Tóth-Czifra. 2020. "Open Science in the Horizon Europe funding programme: what to expect?," in *DARIAH Open*, https://dariahopen.hypotheses.org/968.

[7] See, for instance Erzsébet Tóth-Czifra. 2020. ”The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem”. In Jennifer Edmond (ed.), *Digital Technology and the Practices of Humanities Research*. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0192

- FAIR Data principles are generic and discipline-agnostic principles that have emerged in the context of natural sciences.[8] On a day-to-day level, researchers and end users are looking for data not in the abstract, but always from a very specific disciplinary angle. For example, to properly assess the fitness-for-purpose of a lexicographic dataset, lexicographers and researchers need specific information that goes beyond generic metadata formats and focuses on the specific questions of linguistic scope, structural model, markup granularity etc. FAIR Data on its own is not sufficient for such queries.
- FAIR Data principles focus more on metadata than content[9], which is why the implementation of FAIR Data principles does not always facilitate actual content reusability. In fact, studies have shown that there is a significant gap between the self-identified adherence to the principles of FAIR, and concrete implementations[10].
- FAIR Data principles focus on end products rather than workflows. That is why the more dynamic principles of *provenance* and *stewardship* are needed to account for different life cycles of datasets in order to guarantee their sustainability.

Certified repositories

Stable and trustworthy hosting of datasets has been recognized as a key infrastructural challenge for providing long-term access to various types of data. Community-based initiatives have led to the development and implementation of important core-level certification mechanisms for *data repositories* such as the Data Seal of Approval (2008-2018)[11] and the CoreTrustSeal.[12] Core certification involves a minimally intensive workflow whereby data repositories provide evidence that they are sustainable and trustworthy by conducting an internal self-assessment, which is then peer-reviewed. We have adopted a similar approach in the development of the LexSeal but with a focus on individual lexicographic datasets.

The Heritage Data Reuse Charter

A number of European organizations such as APEF, CLARIN, Europeana, E-RIHS and projects such as Iperion-CH and PARTHENOS joined forces under the leadership of

---

[8] On the dominant impact on STEM on FAIR and the need for domain-specific implementations, see Deniz Beyan, Oya, Chue Hong, Neil, Cozzini et al. (2020). Seven Recommendations for Implementation of FAIR Practice. Zenodo. http://doi.org/10.5281/zenodo.3904140.

[9] See, for example European Commission Expert Group on FAIR Data (2018), Turning FAIR into Reality: Final Report and Action Plan https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf. doi: 10.2777/1524.

[10] For instance, a survey of 100 datasets in journals that publish ecological and evolutionary research with a strong personal digital archiving policy found that 56% of them were incomplete, and 64% were archived in a way that partially or entirely prevented reuse. See Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biol 13(11): e1002295. https://doi.org/10.1371/journal.pbio.1002295.

[11] Developed in 2008 by DANS (Data Archiving and Networked Services) in response to the requirement from two Dutch funding organizations KNAW and NWO to create a seal of approval that would help ensure that archived data can still be found, understood and used in the future. In the first quarter of 2009, the DSA was handed over to an international board. DSA was focused on the certification of data repositories through peer-reviewed self-assessment. Aimed at data providers, funders and end-users, DSA assured the reliability, accessibility and reusability of stored data. See Leeuw, L. 2019. Data Seal of Approval (DSA). DANS. https://doi.org/10.17026/dans-28z-njxq

[12] In 2018, the DSA was merged with the ICSU Word Data System (WDS) into a CoreTrustSeal, an international, community based, non-governmental, and non-profit certification organization promoting sustainable and trustworthy data infrastructures. To manage its finances, CoreTrustSeal is a legal entity under Dutch law (CoreTrustSeal Foundation Statutes and Rules of Procedure) governed by a Standards and Certification Board composed of 12 elected members representing the Assembly of Reviewers. Since 1 February 2018, an administrative fee of EUR 1,000 is charged for the review of a CoreTrustSeal certification. See https://www.coretrustseal.org/apply/administrative-fee/

DARIAH to agree on The Heritage Data Reuse Charter as a set of principles and mechanisms for improving the conditions for the use and re-use of cultural heritage data by researchers.[13] These high-level principles (Reciprocity, Interoperability, Citability, Openness, Stewardship and Trustworthiness) have been selected in order to articulate mutual commitment to clearly articulated conditions of reuse, data sharing formats, standards, processes and protocols, hosting and maintenance responsibilities, as well as the richest possible track of documentation and provenance information. Because data reuse is high on the list of priorities of research infrastructures and projects such as ELEXIS, the proposed Lexical Data Seal of Compliance has been aligned with the principles of the Heritage Data Reuse Charter.

LexSeal in context
LexSeal promotes the use of standards and standard data formats, while acknowledging the reality that not all datasets are indeed standardized and that proper documentation can go a long way in assuring the reuse of all datasets.

LexSeal grounds generic FAIR Principles in the disciplinary realities of lexicographic research and practice. As a domain-specific implementation, LexSeal is more substantial than FAIR, but nonetheless situated in a minimally intensive workflow based on self-assessment questionnaires that should be relatively easy for dataset creators and providers to respond to. A LexSeal-certified lexicographic dataset *is* a FAIR dataset, but the domain-specific information required for LexSeal compliance makes it eminently richer and more useful. Unlike FAIR, LexSeal is a truly community-governed initiative from and for experts who work directly with lexicographic datasets.

The current proposal is a product of ongoing discussions between the DARIAH WG “Lexical Resources”, the ELEXIS Standards Committee and the ELEXIS Integration and Sustainability Committee.

Challenges

Evaluation culture

The challenge of providing and adopting clear evaluation criteria and procedures for digital publications in general have been highlighted both by individual scholars[14] as well as

[13] Erzsébet Tóth-Czifra, Laurent Romary. 2020. The Heritage Data Reuse Charter: from principles to research workflows. ⟨halshs-02475692⟩

[14] See, for instance, Baillot, Anne. 2016. ‘A Certification Model for Digital Scholarly Editions’, October. https://halshs.archives-ouvertes.fr/halshs-01392880; Eve, Martin. 2020. ‘Violins in the Subway: Scarcity Correlations, Evaluative Cultures, and Disciplinary Authority in the Digital Humanities’. In *Digital Technology and the Practices of Humanities Research*, edited by Jennifer Edmonds. Cambridge: Open Book Publishers; Takats, Sean. 2013. ‘A Digital Humanities Tenure Case, Part 2: Letters and Committees’. *Quintessence of Ham*. http://quintessenceofham.org/2013/02/07/a-digital-humanities-tenure-case-part-2-letters-and-committees/; and Zundert, Joris J. van, Smiljana Antonijević, and Tara L. Andrews. 2020. ‘6. “Black Boxes” and True Colour — A Rhetoric of Scholarly Code’. In *Digital Technology and the Practices of Humanities Research*, edited by Jennifer Edmond. Open Book Publishers. https://doi.org/10.11647/OBP.0192.06.

research institutions and scholarly associations.[15] Broadly speaking, the available approaches can be grouped into two broad categories:

- traditional peer-review, exemplified, for instance by RIDE: A Review Journal for Digital Editions and Resources[16] and
- certification systems such as the DINI-Zertifikat for Open Access Publication Services, provided by the Deutsche Initiative für Netzwerkinformation[17] or the already mentioned DataSeal and CoreTrustSeal initiatives.

Each approach comes with its own set of challenges. The traditional pre-publication peer-review solved the problem of the publication and dissemination costs in the age of print: a selection process had to be established because it was too expensive to print everything. But this kind of pass/fail selection process is completely outdated in the digital age: it takes too long, and has questionable implications for innovation.[18] Certification mechanisms, on the other hand, often require significant organizational investment and may not seem at first necessarily appropriate for individual datasets: “it takes more than checking boxes like “TEI-based”, “Open Access” and “Long Time Archiving” to conceive such a certification model for digital scholarly editions in order for it to be truly useful.”[19]

The proposed Lexical Data Seal of Compliance is an attempt to create a community-run certification mechanism based on the peer-review of not datasets themselves, but self-assessment reports provided by their creators. The proposed mechanism is based on shared responsibilities and trust-building among the involved stakeholders. Like standardization efforts in general, certification is also a social construct: for its value to be established, a critical mass of users and endorses will be needed.

---

[15] See for instance, the criteria developed by the Institut für Dokumentologie und Editorik: Patrick Sahle et al. 2014. Criteria for Reviewing Scholarly Digital Editions, version 1.1: http://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/; Ulrike Henny, Frederike Neuber et al. 2017. Criteria for Reviewing Digital Text Collections, version 1.0: https://www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/; Anna-Maria Sichani, Elena Spadini et al. 2018. Criteria for Reviewing Tools and Environments for Digital Scholarly Editing, version 1.0: https://www.i-d-e.de/publikationen/weitereschriften/criteria-tools-version-1/; and the MLA Statement on the Scholarly Editions in the Digital Age. 2016. https://www.mla.org/content/download/52050/1810116/rptCSE16.pdf

[16] RIDE was established in 2014 with the goal of providing a forum “in which expert peers criticise and discuss the efforts of digital editors in order to value their work and also to improve current practices and advance future developments.” In 2017, the focus of the journal expanded to include digital text collections (“digital resources that involve the collecting, structuring and enrichment of textual data from various humanities disciplines such as Literary Studies, Linguistics and History”), and in 2020, once more, to cover “the reviewing of software, particularly of tools and environments for scholarly editing.”

[17] https://dini.de/dienste-projekte/dini-zertifikat/

[18] See Tennant, Jonathan P., Jonathan M. Dugan, Daniel Graziotin, Damien C. Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, et al. 2017. ‘A Multi-Disciplinary Perspective on Emergent and Future Innovations in Peer Review’. F1000Research 6 (November): 1151. https://doi.org/10.12688/f1000research.12037.3; Risam, Roopika. 2014. ‘Rethinking Peer Review in the Age of Digital Humanities’. 2014. https://digitalcommons.salemstate.edu/cgi/viewcontent.cgi?article=1002&context=english_facpub; Neylon, Cameron. 2010. ‘Peer Review: What Is It Good For?’ 2010. https://cameronneylon.net/blog/peer-review-what-is-it-good-for/; Fitzpatrick, Kathleen. 2011. Planned Obsolescence: Publishing, Technology, and the Future of the Academy. New York: New York University Press; and Fyfe, Aileen, Kelly Coate, Stephen Curry, Stuart Lawson, Noah Moxham, and Camilla Mørk Røstvik. 2017. ‘Untangling Academic Publishing: A History of the Relationship between Commercial Interests, Academic Prestige and the Circulation of Research’. Zenodo. https://doi.org/10.5281/zenodo.546100.

[19] Anne Baillot. A certification model for digital scholarly editions: Towards peer review-based data journals in the humanities. In *Digital Scholarly Editing: Theory, Practice, Methods*, Université d'Anvers, Oct 2016, Anvers, Belgium. ⟨halshs-01392880⟩

Competing interests

As a knowledge domain, lexicography is of interest to both the scholarly and the commercial sector.[20] Researchers have a vested interest in having full access to lexical data, while commercial providers have a vested interest in controlling access to their intellectual property. The two interests are in fundamental opposition to one another and impossible to reconcile in every single instance.

Even within the public sector itself, the landscape is varied and full open access has not yet become the established norm in every part of the world. Some publicly funded institutions may provide access to their lexicographic content via graphical user interfaces such as online portals or dedicated applications, but they may still not make the source data available for download or via APIs.

Yet, regardless of the degree of access which the data provider is ready to open up to, all the stakeholders have an interest in having access to clear and transparent information *about* lexicographic datasets. A dataset provider may not wish to grant direct access to the content of a given resource, but would have nothing to lose from granting access to the high-quality metadata about the resource, including but not limited to the questions of possible licensing models. A researcher, on the other hand, may prefer full access, but would also find it useful, in cases when such access is not open and free, to learn about the resource itself and the attached licensing opportunities.

Discussions around this topic included proposals to develop a multi-tiered LexSeal certification system (gold, silver and bronze) which would address the gamut running from closed, commercial datasets to fully open datasets, but this approach was abandoned for practical reasons: a certification system which is based on the hierarchical assessment of openness would not be very attractive to commercial providers, because they would know from the beginning that they could attain only the “lowest” version of the LexSeal. At the same time, limiting certification eligibility to only publicly available datasets would paint an unrealistic picture of the domain.

For LexSeal to be maximally attuned to the realities in which lexicographic datasets are subject to competing interests and legitimately diverse levels of openness, we propose to provide two types of certification:

1. **LexSeal** would certify the information about the resource itself, the degree to which it is well-documented and aligned with best practices; and
2. **LexSeal+** would, in addition, certify the degree to which the resource adheres to the principle of open access and collaborative knowledge creation.

[20] A large portion of dictionaries for general users that are produced today are aimed at the mass market and produced by commercial publishers. On the notion of a “dictionary for general users,” see Henri Béjoint. 2016. “Dictionaries for General Users: History and Development; Current Issues.” In Durkin, Philip, ed. *The Oxford Handbook of Lexicography*. Oxford and New York, NY: Oxford University Press. 7-24. For an attempt to define a scholarly dictionary, see Dirk Kinable. 2015. “Reflections on the concept of a scholarly dictionary.” In *Kernerman Dictionary News,* no. 23. 11-12. https://www.kdictionaries.com/kdn/kdn23_2015.pdf#page=11

Dynamic landscape

The proposed certification system for lexical datasets should be capable of evolving over time. Technologies change, new standards may appear, new business models for providing access to lexical resources may be developed. We do not want LexSeal to solidify the current state of affairs. Instead, we want LexSeal to be able to move with the times.

To address this challenge, we propose that the LexSeal certificates be versioned objects. The first official release of the LexSeal, which we expect to come into being after the official constitution of the Governing Body, will be versioned as LexSeal 1.0. Similar to, for instance, Creative Commons licenses, which have gone through four versions between December 2002 and November 2013, we propose that LexSeal be built, from the ground up, as a versionable system.

The criteria for making minor and major updates shall be left to the future Governing Body.

Required effort

It is very difficult to predict the scope of the institutional infrastructure needed for running and maintaining the LexSeal as a community-based certification system. Can this process be incorporated into the day-to-day business of the governing bodies without much disruption? Will the institutional overhead of maintaining the quality control of the process be manageable? And, finally, will the community buy-in in terms of peer-review volunteer contributions be significant enough to ensure a smooth operation?

To answer these and similar questions, we propose the following mitigating measure: after completing the ELEXIS deliverable but before negotiating a possible formal agreement with DARIAH and/or CLARIN, the ELEXIS team behind the LexSeal will run a trial review process. The process take several steps:

- drafting the full Self-Assessment Questionnaire based on the recommendations expressed in this report (see below);
- inviting members of the DARIAH Working Group “Lexical Resources”, ELEXIS partners and, potentially, ELEXIS Observers to volunteer a small number of lexicographic datasets for trial evaluation;
- inviting members of the DARIAH Working Group “Lexical Resources”, interested ELEXIS partners and, potentially, colleagues from the ELEXIS Observer Organizations to participate in a trial review process as reviewers.
- conducting a trial review process; and
- evaluating the process once it has been completed.

The trial period will be used to gain real-life experience and identify possible shortcomings in both the workflow *and* the drafted topics for the self-assessment questionnaire. The institutions volunteering for the trial period shall be made aware that no Lexical Seal will be awarded at the end of the trial period and that the future Governing Board may require that they resubmit or update their self-assessment questionnaires, should LexSeal become an officially recognized certification mechanism.

Certification

The Lexicographic Data Seal of Compliance is established through a three-stage process:

- a self-assessment questionnaire by creators and providers of lexicographic datasets;
- a review of the self-assessment questionnaire by members of the LexSeal Assembly of Reviewers
- approval of the LexSeal review by the Governing Board

Levels

We propose two levels of certification: LexSeal and LexSeal+.

LexSeal is awarded to lexicographic datasets which meet the LexSeal *baseline criteria of compliance* in every LexSeal assessment category.

LexSeal+ is awarded to lexicographic datasets which, in addition to meeting the baseline criteria for compliance in every LexSeal assessment category, also meet the *advanced criteria of compliance* in three out of five LexSeal assessment categories and at least one LexSeal+ assessment category.

LexSeal assessment categories are:

- Trustworthiness
- Interoperability
- Stewardship
- Citability

LexSeal+ assessment categories are:

- Reciprocity
- Openness

A lexicographic dataset can be checked against the baseline and advanced criteria of compliance in each assessment category, with the exception of *trustworthiness*, which only has one required set of baseline criteria that all LexSeal-certified datasets must meet.

Each answer provided in the self-assessment questionnaire is assessed as satisfactory or unsatisfactory. A reviewer will mark the dataset as compliant in a given category if he or she considers all the answers in the given category and at a given criterium level as satisfactory.

Self-assessment

LexSeal Assessment Categories

TRUSTWORTHINESS	
The trustworthiness of the lexicographic resource is assessed by the degree to which the given dataset is documented in terms of its scope, content, authorship and provenance.	
Baseline Criteria	*Advanced Criteria*
<u>Expression of scope</u> • Describe the resource and its intended audience(s). What can the resource be used for? <u>Linguistic content</u> • What are the source/target language(s)? • What language periods are covered (historical, contemporary)? • What language varieties are covered (in terms of standard languages, geographic varieties, sociocultural registers etc.) <u>Bibliographic information</u> • Who created, edited and published the lexicographic dataset (authors, editors, other contributors, publishers, year of publication(s)? • Does the lexicographic dataset exist in multiple versions? <u>Provenance</u> • What is the origin of the lexicographic dataset (manuscript, print, born-digital)? • Who is the owner/rights holder of the dataset, and, if applicable, the source from which the dataset was generated? • What quality control measures were taken to assure the integrity of the text (for instance: if the text has been OCR'ed, how was it corrected?)	N.A.

INTEROPERABILITY

The interoperability of the lexicographic dataset is assessed by the degree to which a given dataset is documented in terms of its structure, format, reusability and licensing

Baseline Criteria	*Advanced Criteria*
Text format • Is the text machine readable? Data model • Is the underlying data model semasiological or onomasiological? Reusability • Is explicit, human and machine-readable licensing information provided? • Are there any technical prerequisites for using the dataset (specific tools, databases, fonts etc.)?	Encoding • Is the dataset semantically encoded? At what level of granularity (lemmas, senses, definitions etc.) • Is the encoding aligned with de facto and de jure standards (Unicode, ISO, W3C, TEI etc.) • Is a schema made available with the dataset to check its validity?

STEWARDSHIP

The stewardship of a given lexicographic dataset is assessed in terms of the commitment to stable curation and hosting of the lexicographic dataset.

Baseline Criteria	*Advanced Criteria*
• Is somebody in charge of curating the lexicographic dataset? • Is the information provided in this questionnaire available to the users as part of explicit metadata or in some other form of documentation (for instance, in a README file)? • Is there stable institutional hosting and provisions for the long-term availability of the dataset?	• Is the dataset hosted in a CoreTrustSeal compliant repository

CITABILITY

The citability of a lexicographic dataset is assessed in terms of the degree to which the user is made aware of the preferred way of referring to the dataset.

Baseline Criteria	*Advanced Criteria*
• Is there a clear statement on how to refer to the dataset in a human-legible fashion (in scholarly publications, for example)?	• Is there a clear statement on how to refer to parts of the dataset (for instance, an individual entry, or an individual sense) in a human-legible fashion? • Are there technical means to refer to the resource or parts thereof (PIDs, querying mechanisms etc.)?

LexSeal+ Assessment Categories

OPENNESS

The openness of a lexicographic dataset of a lexicographic dataset is assessed in terms of the provision of free, unrestricted access to the content.

Baseline Criteria	*Advanced Criteria*
• Are parts of the lexicographic dataset available for free, unrestricted access, download and reuse (for instance, a lemma list, or partial/specific entry components such as alternative spellings, phonetic transcriptions, sense divisions, synonyms/antonyms, usage information, definitions, examples, translation equivalents, MWEs etc.)?	• Is the whole lexicographic dataset available for free, unrestricted access, download and reuse?

RECIPROCITY

The reciprocity of a lexicographic dataset is assessed in terms of the provider's openness to receive feedback from the users.

Baseline Criteria	*Advanced Criteria*
• Is there a contact person that the user can submit questions about the resource to?	• Is there a possibility for users to submit suggestions for improvement of the resource (corrections, enrichment etc.)?

Review process

Reviewers are members of the Assembly of Reviewers, which is not a fixed scientific committee but a fluid body, similar to the pools of reviewers that review scholarly articles or conference submissions.

The initial pool or reviewers are self-nominated and approved by the LexSeal Governing Body. A call for reviewers can be made to the wider lexicographic community via mailing lists (incl. but not limited to CLARIN, DARIAH, ELEXIS, incl. ELEXIS Observers).
The Assembly of Reviewers shall be recruited dynamically: representatives of the reviewed lexicographic data providers are subsequently invited to become reviewers themselves.

Reviewers receive a certificate of recognition for their role in reviewing datasets, which they can mention on their CVs the same way that they highlight the journals or conferences that they write reviews for. Members of the Assembly of Reviewers can grant permission for their names to be listed on the LexSeal website.

Submissions shall be managed through an instance of Sciencesconf.org or a similar platform hosted by the DARIAH WG “Lexical Resources.”

Resubmissions shall be allowed if:

- a previously evaluated resource has undergone substantial editorial changes which merit a new review (and potentially a “better” evaluation)
- a resource has been rejected and the authors can document concrete improvements to the resource and its documentation since the last evaluation

Governance

The LexSeal shall be administratively anchored in the DARIAH WG “Lexical Resources” to ensure sustainability beyond the end of ELEXIS as a funded project.

We propose for LexSeal to be overseen by the LexSeal Governing Board, consisting of:

- 2 members from DARIAH, nominated by the DARIAH Board of Directors
- 2 members from CLARIN, nominated by the CLARIN Board of Directors
- 2 members from the ELEXIS legal entity which is expected to be formed in 2022

The role of the LexSeal Governing Board shall be:

- to vouch for the integrity of the review process
- to appoint reviewers
- to approve completed reviews
- potentially, to adopt possible changes to the procedures
- potentially, to prepare new versions of the LexSeal and offer them to the community for public consultations
- potentially, to officially adopt new versions of the LexSeal after a round of public consultations

New versions of the LexSeal will not be frequent, so it is unlikely that serving on the LexSeal Governing Board will consume an exorbitant amount of time. At the same time, we shouldn't minimize the effort required. The proposed trial review period (see above) will help us understand better the extent to which the Governing Board will need to be consulted in the review process.

Future perspectives

Short term (2021-2022)

- Certification trial run after the submission of the ELEXIS deliverable
- Official proposal to DARIAH and CLARIN
- Establishment of the Governing Body
- First certificates being awarded

Medium term (2022-2025)

- Analysis of the community uptake
- Infrastructural consolidation: discussions on the possibility of building a registry of LexSeal certified resources
- Explore integration potential with SSHOC Marketplace, the TRIPLE discovery platform and, generally, EOSC
- Scanning the landscape for possible project funding opportunities for implementing such integrations

Long term (2025 and beyond)

- Depending on the outcomes of the community uptake analysis, the assessment of the opportunities for infrastructural consolidation, and the identification of possible funding sources, work toward robust, technologically innovative solutions for disseminating information about LexSeal-certified datasets; and providing federated access to LexSeal+ certified datasets.