

CREATING A UNIVERSAL INFORMATION LANGUAGE

OSCAR BENDER-STONE

ABSTRACT. Welkin is a formalized programming language to store information. We introduce its use cases and rigorously define its syntax and semantics. From there, we introduce the bootstrap, making Welkin completely self-contained under a meta-theory based on combinators, equivalent to a provably minimal fragment of arithmetic.

CONTENTS

1. Introduction	1
1.1. Goals	2
1.2. Organization	3
2. Motivating Example	3
3. Syntax	3
4. Semantics	4
4.1. ASTs	4
4.2. Representations	4
4.3. Universal Systems	6
5. Information Organization	6
5.1. Impossible Classes	7
5.2. Efficient Querying	7
6. Bootstrap	7
7. Conclusion	7
References	8

1. INTRODUCTION

Information Management (IM) is an open area of research as a result of the depth and breadth of disciplines. In terms of depth, many areas are often specialized, requiring an immense understanding of the broader concepts involved and nomenclature used. This specialization is evident in the sciences, as explored in [Fan13], [CF14]. Additionally, in terms of breadth, creating common representations shared across sub-disciplines can be difficult. For example, mathematics has extremely diverse disciplines, and connecting these areas is an open problem in scalability [Car+21]. Moreover, creating a standardized form across communities is challenging. In other subjects, like the social sciences, there are no standard terms [Arc+06], and in the humanities, representing certain artifacts as data is involved [Har+20]. More broadly, IM *itself* is divided from distinct approaches that lack interoperability [AJ23]. Certain frameworks equate IM to Knowledge Management (KM) and assert that information must be true [Edw22]. These problems, in both faithfully and broadly storing information, demonstrate the enormous task of effective IM.

In response to these challenges, several solutions have been proposed, but none have been fully successful. In the sciences, a group of researchers created the Findable Accessible Interoperable Resuable (FAIR) guidelines [Wil+16]. Instead of providing a concrete specification or implementation, FAIR provides best practices for storing scientific information. However, multiple papers have outlined problems with these overarching principles, including missing checks on data quality [Gui+25], missing

expressiveness for ethics frameworks [Car+21], and severe ambiguities that affect implementations [Jac+20]. Along with the sciences, there are several projects for storing mathematical information (see [CF09] for more details). Older proposals, including the QED Manifesto [KR16] and the Module system for Mathematical Theories (MMT), aimed to be more general and have seen limited success. More centralized systems, like mathlib in the Lean proof assistant [Com20], have seen adoption but do not give equal coverage nor are interoperable with other systems. Beyond more “hard” fields, IM in the humanities has few models, including an adaption of FAIR [Har+20] and discipline specific, linked databases in the PARNTHEOS project [Hed+19]. Each of these proposals, even within specific fields, fail to accommodate for all of the mentioned challenges.

In addition to domain specific proposals, there are approaches for general IM which still fail to resolve all issues. One prominent example is Burgin’s theory of information [Bur09] that comprehensively includes many separate areas for IM, including the complexity-based Algorithmic Information Theory (AIT), through a free parameter called an “infological system”, which encompasses domain specific terminology and concepts. In contrast to other approaches, Burgin’s generalized theory is flexible and enables greater coverage of different kinds of information [Mik23]. Despite this coverage Burgin does not closely tie the free parameter with his formal analysis of AIT, making it unclear how to use this in a practical implementation. Broad frameworks for IM, along with the specific proposals, have severe shortcomings, highlighting major obstacles for IM.

This thesis introduces a language to resolve these issues. I call this language **Welkin**, based on an old German word meaning cloud [Dic25]. The core result of this thesis is proving that Welkin satisfies three goals: is **universal**, **scalable**, and **standardized**. For details, see Section 1.1. The core idea is to generalize Burgin’s free parameter and enable arbitrary representations in the theory, controlled by a computable system. The notion of representation builds on Peirce’s semiotics, or the study of the relationship between a symbol, the object it represents, and the interpreter or interpretation that provides it that meaning [Atk23]. Moreover, to address queries on the validity of truth, we use a relative notion that includes a context managed by a formal system. Truth can then be determined on an individual basis, providing flexibility to any discipline. The focus then shifts to the usefulness of representations based on a topological notion of “collapse”, which we call **coherency**. This approach is inspired by coherentism, a philosophical position that states truth is determined in comparison to other truths. [Bra23]. We incorporate ideas from coherentism to identify which representations identify their corresponding objects, and we define information as an invariant under these coherent representations. We include definitions on a *working* basis as what is most practical, not an epistemological stance that can be further clarified in truth systems. Additionally, we keep the theory as simple as possible to make scalability and standardization straight-forward.

1.1. Goals.

- **Goal 1: Universality.** The language must include unspecified, user created parameters to accomodate for arbitrary concepts and ideas.

- **Goal 2: Scalability.** The database must appropriately scale to broad representations of information. Local queries must be efficient. Certificates must be available to prove cases where optimal representations have been achieved.
- **Goal 3: Standardization.** The language needs a rigorous and formal specification. Moreover, the bootstrap must be formalized, as well as an abstract machine model. The grammar and bootstrap must be fixed to ensure complete forwards and backwards compatibility.

1.2. Organization.

This thesis is organized as follows.

- **Section 2: Motivating Example** **Section 2.** Introduces a high-level example to explain the core concepts in Welkin.
- **Section 3: Syntax** **Section 3.** Provides the syntax and proves the language is unambiguous.
- **Section 4: Semantics** **Section 4.** defines the interpretation on the AST and representations. Also presents coherency as a central notion in the theory, and provides a working definition of information.
- **Section 5: Information Organization.** **Section 5** Develops the optimal information system based on a (localized) Greedy algorithm. Additionally, creates certificates to demonstrate when a given representation is locally optimal.
- **Section 6: Bootstrap.** **Section 6** Bootstraps Welkin by proving that there is a Welkin node that contains enough information about the standard. Fulfills Goal 2 with both the Standard AND the complete bootstrap.
- **Section 7: Conclusion** Reviews the work done in the previous sections and explores several possible applications.

2. MOTIVATING EXAMPLE

3. SYNTAX

To keep this thesis self-contained, all recursive definitions are included. For simplicity, we will use the notation a_0, \dots, a_n for a finite list of items. We will revisit the notion “finite” more rigorously in Section 6.

Definition 3.0. The **alphabet of binary words** is $\mathcal{A}_{\text{bword}} ::= \text{bit} \mid . \mid w$, where $\text{bit} ::= 0 \mid 1$. A **binary word** is defined recursively: the symbols 0 or 1 are strings, or if w is a string, then so are $w.0$ and $w.1$. We set concatenation to be right-associative, i.e., $(w.w').w'' = w.(w'.w'')$, and safely abbreviate $w.w'$ as ww' . We write $w \in \text{bword}$ to denote that w is a binary word.

For simplicity, we extend the alphabet to include two common bases: decimal and hexadecimal.

Now, the base encoding for Welkin is in US-ASCII, formally defined below.

Definition 3.1. The **alphabet of words** $\mathcal{A}_{\text{word}} := \mathcal{A}_{\text{bword}} + \mathcal{A}_{\text{dec}} + \mathcal{A}_{\text{hex}}$, where:

- $\mathcal{A}_{\text{dec}} := 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9$
- $\mathcal{A}_{\text{hex}} := A | B | C | D | E | F$

A **word** is a concatenation of either only binary digits, only decimal digits, or hexadecimal digits. Each of these are denoted with different prefixes: decimal has none, binary uses $0b$, and hexadecimal uses $0x$.

Using binary words simplifies the bootstrap, so while these digits are included, they are *defined* in terms of binary words, see Section 6.

Now, the base encoding for Welkin is in US-ASCII, formally defined below.

Definition 3.2. US-ASCII consists of 256 symbols, listed in Table ?.

We reserve the term **string** when a word is explicitly enclosed in delimiters, namely single or double quotes.

Now, we formalize an unambiguous form of EBNF for our use case.

Definition 3.3. BNF consists of productions. Writing $r := a_1 | \dots | a_n$ is shorthand for including the rules $r := a_1, \dots, r := a_n$.

EBNF additionally contains the following abbreviations:

- $r := a_1 | \dots | a_n$ is shorthand for including the rules $r := a_1, \dots, r := a_n$.

Welkin's grammar is displayed in ...

Definition 3.4. A grammar is LL(1) if ...

Theorem 3.5. Welkin's grammar is LL(1). Hence, this grammar is unambiguous.

4. SEMANTICS

4.1. ASTs.

- Semantics on ASTs
 - Terms: graphs
 - For ease of use, include a null node that is the root of the tree. This represents the module itself.
- For information organization: integrate with previous section
 - Emphasize how this is a useful tool and can ensure **new** information content is being created (at least, that can be distinguished from the current module). If already existing, but that doesn't match the user's expectations, they need to refine it! OR, maybe it **does** match similarly with something else! (e.g., hidden connections between math and music)
- Emphasize pragmatics as well, via units

4.2. Representations.

Now we develop the formal framework to discuss information in terms of units, enabling a complete mechanization of Welkin's meta-theory. To keep this section self-contained, we explicitly provide all recursive definitions.

Definition 4.2.0. The **alphabet of units** is $\mathcal{A}_{\text{unit}} = u | \mathcal{A}_{\text{word}}$. A **unit ID** is combination of symbols u_w , where $w \in \text{word}$.

Definition 4.2.1. A **free parameter** is a parameter given an associated ID. No further restrictions are imposed.

We now define representations recursively, using unit IDs and free parameters as the base case.

Definition 4.2.2. Units are recursively defined:

- **Base case:** IDs and free parameters are units.
- **Recursive step:**
 - **Parts:** if u_1, \dots, u_n are finitely many units, then so is their combination $\{u_1, \dots, u_n\}$. A combination defined without a provided ID is called an **anonymous unit**.
 - **Representations:** If u, w, v are units, so is $v \xrightarrow{u} u$. We say v **represents** u . or conversely, u is **represented by** v .

Key equalities:

- $u.\{\} = u$. Acts as a sort of * operator from other languages.
 - To use one level up: $.u$
- $(u \xrightarrow{v} w) \in x \Leftrightarrow x(u) \xrightarrow{x(v)} x(w)$, where $x(u)$ is $x.u$ if $u \in x$ or u otherwise.

Example 4.2.3. Consider a house with a dog, a cat, and a person. We can represent the house as unit house, the dog as unit dog, the cat by unit cat, and the person by unit person. In our Welkin file, we add, `house { dog, cat, person}`. The person has an internal concept of pet and uses it to represent both the dog and cat, which we write as `person { animal --> dog, animal --> cat}`, under the scope of house.

Parts of units are denoted as $u.u'$. Scoping is included to provide namespaces. Moreover, parts enable **interpretations**. We write $u \xrightarrow{v} u'$ in case $u, v, (u \rightarrow v) \in u'$, so u represents v **via** u' . In this case, we say u' is a **context** to $u \rightarrow v$. Note that unlabeled representations can have multiple contexts.

Example 4.2.4. Consider the recursive definition of a binary tree: either it is a null (leaf) node, or it contains two nodes, left and right. We can model this as follows:

- First, create units for each of the notions: `tree {null, left, right}`.
- Next, we write, `tree { null --> .tree, .tree --> left, .tree --> right, {.left, .right} --> .tree}`. Notice that we refer to the *namespace*, thereby enabling recursion. By our scoping rules, writing `tree` would be a *new unit*.
- To impose that the left subtree is *distinct* from the right one, we can use symbols.

An important idea in this example is that the abstraction could be defined *first*, or a concrete model could. For this reason, the choice of how entities are represented is flexible.

Definition 4.2.5. A unit u is **non-trivial** if it is non-empty and does not contain all relations. A unit u is **coherent relative to a context** u' if $u + u'$, the union of these units, is non-trivial.

Remark 4.2.6. This definition is a natural generalization of consistency in first-order logic. We will frequently rely on this result throughout the thesis.

Theorem 4.2.7. *A representation is preserves information modulo \equiv iff the representation modulo \equiv is coherent.*

Remark 4.2.8. This theorem enables truth management via specific contexts, specified as units. The task of finding core truths is then free, left open to flexibility accommodate for any truth management.

4.3. Universal Systems.

Inspired by [Mes12], we prove that scoping is strictly more expressive than without.

Lemma 4.3.9. *Representations with interpretations are undefinable in terms of unlabeled representations.*

Proof. It suffices to note that representing partial computable functions requires combinations. But every transformation under unlabeled representations does not preserve these conditions, hence, representations with interpretations are not definable. \square

Theorem 4.3.10. *Any computable function and its trace under a given string can be represented by units. (TODO: make more precise.)*

TODO: TODO:
define the
generalization to
Padoa's Method
clearer.

Proof. We prove there is a natural embedding from the lambda calculus into our meta-theory, using two parts:

- $\lambda x.f$ is represented as $\{x \rightarrow f\}$.
- $f.g$ is represented as the combination of units, or $\{f, g\}$.

Thus, any λ -term can be represented in te meta-theory, completing the proof. \square

Note that there are multiple ways to prove Theorem 4.3.4, infinitely in fact. This motivates the following definition.

Definition 4.3.11. A universal representation system (URS) is a unit that can represent any representation.

Theorem 4.3.12. *A unit is a universal representation system if and only if it can represent any partial computable function. Moreover, any universal representation system can represent any universal representation system. In particular, representing itself is called reflection.*

TODO: Make this
more precise and
complete proof.

The term *universal* is specifically for expressing *representations* symbolically. The free parameter still needs to be included and is an additional feature on top of partial comptuable functions. However, the *management* of these symbols is done entirely with partial computable functions.

The next section discusses the issue of *managing* the infinitely many choices for URSs.

5. INFORMATION ORGANIZATION

- Main question: **which** universal system to choose? Is this practical?
 - What is a suitable criterion for a base theory?
 - Recall aim: want to mechanically store systems for a database
 - * What if possible performance degredation? Will we get stuck if we start with one architecture? Will we have to adjust later?
 - * Aim is to ensure architecture is completely flexible and can automatically adapt

- * One key metric: ability to store as many systems coherently as possible,
i.e., store as much information as possible
- Main problem: Blum's speedup theorem
 - * Briefly generalize this for slate logic
 - * Show that no single way to completely organize systems based on a computable metric.

This is part of the need for new search techniques!

- * Want to separate search from storage though, but we want to improve

stored results **with** new results. This forms the idea behind the database architecture: have a simple way to store results that automatically gets better with new techniques/results.

- * Need explicit proofs for this! Not sure how to store certificates...

5.1. Impossible Classes.

The reason to restrict our transformations is two-fold. First, we need to ensure we can *verify* them efficiently. Determining whether a morphism between two formal systems exist can be reduced to the Halting problem, and is therefore not practical for defining an optimal formal system. Second, if we include those transformations that we *can* effectively check, no optimal formal system exists.

Theorem 5.1.0. *With respect to the class of all computable transformations that can be computably verified, there is no optimal formal system.*

5.2. Efficient Querying.

Instead of making proofs most efficient as is, we want to support finding optimal representations. But we want to do this from an efficiently queryable system, which *is* the most optimal.

6. BOOTSTRAP

- Provided in bootstrap.welkin file or similar. Main module is welkin, which needs to:
 - Provide slates, so inductive definition of binary strings + variables
 - Explain combinators
 - Explain the basics of universal systems. This is very important!
 - * Can elide proofs IF there is enough information content.
 - * TODO: figure out how proofs might work in this setting
 - Provides syntax and semantics of Welkin itself
 - This thesis: will prove that the AST generated from this file is correct AND that it does, with a suitable interpretation bootstrap itself.
 - Explain how slates expand the envelope for implementations, BUT ensures that the final product, the syntax + semantics checkers and the information organization, can be externally seen!

7. CONCLUSION

- Review of thesis
 - Developed slate logic + bi-translation with FOL

- Developed locally optimal organizational technique that can improve based on annotations/certificates
 - Introduced the language, with a straightforward graph syntax and semantics
 - Builds upon the last section
 - Bootstrapped standard + used coherency condition
- Significance
 - Backwards AND forwards compatible standard that bootstraps itself. Easy for implementations!
 - Applications to any human subject
 - * Sciences
 - * Liberal arts
 - * Economics
 - * Etc.
- Future work
 - Programming language semantics + synthesis
 - * Incorporate broader aspects + intent of users! ESSENTIAL for new programming languages to be able to discuss pragmatics in some way!
 - * Also reproducible AND executable specifications, though creating an engine to execute these is far beyond the scope of the thesis
 - Organizing large corpuses of human text
 - Numerous applications to AI and improving results
 - * Emphasize role of symbol grounding problem in AI

REFERENCES

- [Arc+06] Archambault, Éric, Vignola-Gagné, Étienne, Côté, Grégoire, Larivière, Vincent, and Gingras, Yves, Benchmarking scientific output in the social sciences and humanities: The limits of existing databases, *Scientometrics* **68** (2006) 329–342.
- [Atk23] Atkin, Albert, Peirce's Theory of Signs, Spring2023 ed., Metaphysics Research Lab, Stanford University, 2023.
- [AJ23] Auth, Gunnar, and Jokisch, Oliver, A systematic mapping study of standards and frameworks for information management in the digital era, *Online Journal of Applied Knowledge Management* **11** (2023) 1–13.
- [Bra23] Bradley, F. H., The Principles of Logic, *Mind* **32** no. 127 (1923) 352–356.
- [Bur09] Burgin, Mark, *Theory of Information*, World Scientific, 2009.
- [CF09] Carette, Jacques, and Farmer, William M., A Review of Mathematical Knowledge Management, in *Intelligent Computer Mathematics* (Carette, Jacques, Dixon, Lucas, Coen, Claudio Sacerdoti, and Watt, Stephen M. (eds.), Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 233–246.
- [Car+21] Carette, Jacques, Farmer, William M., Kohlhase, Michael, and Rabe, Florian, Big Math and the One-Brain Barrier: The Tetrapod Model of Mathematical Knowledge, *The Mathematical Intelligencer* **43** no. 1 (2021) 78–87.
- [Car+21] Carroll, Stephanie Russo, Herczog, Edit, Hudson, Maui, Russell, Keith, and Stall, Shelley, Operationalizing the CARE and FAIR Principles for Indigenous Data Futures, *Scientific Data* **8** no. 1 (2021) 108.
- [CF14] Casadevall, Arturo, and Fang, Ferric C., Specialized Science, *Infection and Immunity* **82** no. 4 (2014) 1355–1360.

- [Com20] Community, The mathlib, The lean mathematical library, in *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, Association for Computing Machinery, New Orleans, LA, USA, pp. 367–381.
- [Dic25] Dictionary, Oxford English, welkin, n., Oxford University Press, 2025.
- [Edw22] Edwards, John S., Where knowledge management and information management meet: Research directions, *International Journal of Information Management* **63** (2022) 102458.
- [Fan13] Fanelli, Wolfgang, Daniele AND Glänzel, Bibliometric Evidence for a Hierarchy of the Sciences, *Plos One* **8** no. 6 (2013) 1–11.
- [Gui+25] Guillen-Aguinaga, Miriam, Aguinaga-Ontoso, Enrique, Guillen-Aguinaga, Laura, Guillen-Grima, Francisco, and Aguinaga-Ontoso, Ines, Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles, *Data* **10** no. 12 (2025).
- [Har+20] Harrower, Natalie, Quinn, Mary, Tóth-Czifra, Erzsébet, and others, *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA E-Humanities Working Group*, Berlin, 2020.
- [Hed+19] Hedges, Mark, Stuart, David, Tzedopoulos, George, Bassett, Sheena, Garnett, Vicky, Giacomini, Roberta, and Sanesi, Maurizio, *Digital Humanities Foresight: The future impact of digital methods, technologies and infrastructures*, GOEDOC, Dokumenten-und Publikationsserver der Georg-August-Universität Göttingen, 2019.
- [Jac+20] Jacobsen, Annika, Miranda Azevedo, Ricardo de, Juty, Nick, Batista, Dominique, Coles, Simon, Cornet, Ronald, Courtot, Mélanie, Crosas, Mercè, Dumontier, Michel, Evelo, Chris T., Goble, Carole, Guizzardi, Giancarlo, Hansen, Karsten Kryger, Hasnain, Ali, Hettne, Kristina, Heringa, Jaap, Hooft, Rob W.W., Imming, Melanie, Jeffery, Keith G., Kaliyaperumal, Rajaram, Kersloot, Martijn G., Kirkpatrick, Christine R., Kuhn, Tobias, Labastida, Ignasi, Magagna, Barbara, McQuilton, Peter, Meyers, Natalie, Montesanti, Annalisa, Reisen, Mirjam van, Rocca-Serra, Philippe, Pergl, Robert, Sansone, Susanna-Assunta, Silva Santos, Luiz Olavo Bonino da, Schneider, Juliane, Strawn, George, Thompson, Mark, Waagmeester, Andra, Weigel, Tobias, Wilkinson, Mark D., Willighagen, Egon L., Wittenburg, Peter, Roos, Marco, Mons, Barend, and Schultes, Erik, FAIR Principles: Interpretations and Implementation Considerations, *Data Intelligence* **2** nos. 1–2 (2020) 10–29.
- [KR16] Kohlhase, Michael, and Rabe, Florian, QED Reloaded: Towards a Pluralistic Formal Library of Mathematical Knowledge, *Journal of Formalized Reasoning* **9** (2016) 201–234.
- [Mes12] Meseguer, José, Twenty years of rewriting logic, *The Journal of Logic and Algebraic Programming* **81** no. 7 (2012) 721–781.
- [Mik23] Mikkilineni, Rao, Mark Burgin's Legacy: The General Theory of Information, the Digital Genome, and the Future of Machine Intelligence, *Philosophies* **8** no. 6 (2023).
- [Wil+16] Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand Jan, Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas, Boiten, Jan-Willem, Silva Santos, Luiz Olavo Bonino da, Bourne, Philip E., Bouwman, Jildau, Brookes, Anthony J., Clark, Tim, Crosas, Mercè, Dillo, Ingrid, Dumon, Olivier, Edmunds, Scott C., Evelo, Chris T. A., Finkers, Richard, González-Beltrán, Alejandra N., Gray, Alasdair J. G., Groth, Paul, Goble, Carole A., Grethe, Jeffrey S., Heringa, Jaap, Hoen, Peter A. C. 't, Hooft, Rob W. W., Kuhn, Tobias, Kok, Ruben G., Kok, Joost N., Lusher, Scott J., Martone, Maryann E., Mons, Albert, Packer, Abel Laerte, Persson, Bengt, Rocca-Serra, Philippe, Roos, Marco, Schaik, Rene C. van, Sansone, Susanna-Assunta, Schultes, Erik Anthony, Sengstag, Thierry, Slater, Ted, Strawn, George O., Swertz, Morris A., Thompson, Mark, Lei, Johan van der, Mulligen, Erik M. van, Velterop, Jan, Waagmeester, Andra, Wittenburg, Peter, Wolstencroft, Katy, Zhao, Jun, and Mons, Barend, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3** (2016).