# A Flexiformal Model of Knowledge Dissemination and Aggregation in Mathematics

Mihnea Iancu, Michael Kohlhase

Computer Science, Jacobs University, Bremen, Germany
initial.last@jacobs-university.de

**Abstract.** In the traditional knowledge dissemination process in mathematics and sciences, authors write semi-selfcontained articles which are then published in journals, conference proceedings, preprint archives, and/or given as talks. Other scientists read these, extract the new knowledge, integrate it into their personal mental model of the field, and use this as the basis for creating new knowledge which is disseminated in the same form.
Somewhat surprisingly, this process has not been modeled from a formal or content-based perspective even though it is at the heart of human MKM and DML.
In this paper we tackle this problem starting from the practice of beginning papers with a "recap", which briefly introduces context, terminology, and notations and thus ties the paper into the knowledge commons. We propose a flexiformal model for knowledge dissemination and its aggregation into a communal, shared knowledge commons based on theory graphs and the newly introduced realms.

## 1 Introduction

Global mathematical knowledge grows – at least – at a rate 120,000 published articles a year to a current crop of about 3.5 Million articles. Even though these articles are scattered over several thousand journals they – together with papers in conferences, preprints in online or local archives, and talks given in seminars – function as a coherent scientific commons of communal knowledge about the various domains of mathematics. Other scientists read these documents, extract the new knowledge, integrate this into their personal mental model of the domain, and use this as the basis for creating new knowledge. This, in turn, is disseminated again through articles, conference papers, preprints, and talks, itself contributing to the knowledge commons.

In this process of knowledge dissemination and aggregation, scientific documents (articles, papers, preprints, and talks) play a great role: they have evolved from printed pamphlets or books and from postal letters in which a mathematician described progress to a colleague – and were then passed around by the latter among colleagues. Documents are assembled into topical journals and conference proceedings volumes, which are in turn assembled into libraries (physical and

virtual ones), which give researchers and practitioners access to the scientific document commons – modulo physical distribution methods like inter-library loan and access right restrictions like membership or commercial constraints. Documents are even classified into a domain-based classification schemes like the Math Subject Classification (MSC), and disseminated in information systems like Math Reviews and Zentralblatt Math.

Today's mathematical documents have a specific conventionalized structure and metadata which not only supports the production/dissemination processes outlined above, but also – we claim – the individual and communal aggregation processes which turn the document collections into a (virtual) knowledge space which mathematicians can operate on to find and apply existing knowledge and create new insights and knowledge.

In formalized mathematics, the situation is very different. Even though collections of formalized mathematics call themselves "libraries", the concept of a "formal document" does not exist or degenerates to a "file" which contains the formal development and possibly includes other files. Explanations for humans are generally relegated to comments or the informal literature described above (publishing about formalizations).

Notable exceptions are the Mizar Mathematical Libary [MizLib] and the Mizar-inspired ISAR format in Isabelle [Wen07]. Both of these contain enough information to generate conventionally structured documents for publication,. e.g. Mizar articles in the Journal of Formalized Mathematics [JFM]. Dissemination, quality control, and "marketing" of results and developments is usually ad-hoc in formalized mathematics. Aggregation of developments into a knowledge space is ephemeral and executed by loading files with formal developments into the memory of a theorem prover or proof checker.

On the other hand, libraries of formalized mathematics directly represent the structure of a mathematical knowledge commons, usually in graph of files and file inclusions or a graph of theories and theory morphisms (see [RK13a] for a survey). The respective graphs supply identifiers for knowledge items and detail their relations to each other.

It stands to reason that the two dissemination and aggregation approaches can profit from each other. The scientific publication process can profit from a more explicitly represented knowledge commons, which enables added-value services for finding, understanding, and applying relevant knowledge items – after all the document/knowledge space even in mathematics is much too large and complex for a single human to process. Of course a prerequisite for this is computer support in the aggregation of the knowledge space. Conversely, formal libraries can profit from a dissemination process based on the publication of self-contained documents to scale the secondary aspects (quality control, checkpointing, citation stability, persistence, attention management) of assembling large bodies of knowledge. Even though formal developments are machine-checkable, their authoring, maintenance, refactoring, ... are processes that need at least some human intervention.

To reap these benefits we need a joint generalization of the two approaches to dissemination and aggregation that combines their advantages. But before we design such a system, we need a content-oriented model of the informal publication process. Somewhat surprisingly, such a model does not exist, even though knowledge dissemination and next-generation publication systems are a the heart of MKM and DML.

In this paper we propose a content-oriented model for knowledge dissemination and its aggregation into a communal, shared knowledge commons. As we make use of our previous development of the *flexiformal* – i.e. supporting flexible degrees of formality [Koh13] – OMDoc format [Koh06], which can represent formal and informal mathematical documents and developments, we think of this as a flexiformal model.

We use the practice of starting mathematical documents with a "recap", which briefly introduces context, terminology, and notations and thus ties the paper into the knowledge commons as a starting point and model it based on OMDoc/MMT theory graphs and the newly introduced realms [CFK14].

In Section 2 we briefly review the structure of mathematical documents and build our intuitions about "recaps" by looking at some examples. We discuss how to represent them using theory graphs in Section 3. Section 4 concludes the paper and discusses future work.

## 2 Common Ground in Mathematical Documents

With **dissemination** we mean the process of assembling a mathematical document for the purpose of publication. We use the term **aggregation** for the process of an individual integrating the knowledge gained from reading or experiencing the respective document into their mental model of the domain. For now we will use these two concepts intuitively only, it is the purpose of this paper to propose a more rigorous model for them. As a first step, we will now have a closer look at the practices in formal and informal mathematics.

### 2.1 The Structure of Informal Mathematical Documents

Mathematical documents traditionally have:

1. A **front/backmatter** and **page margins**, which identify the scientific metadata: *i*) author's names, affiliations, and addresses, *ii*) publication venue, date, and fragment identifiers (e.g. page numbers), *iii*) classification data, e.g. keywords or MSC codes, *iv*) acknowledgements of contributions of other researchers or funding agencies. *v*) access conditions, e.g. copyright, confidentiality designations, or licenses.
2. An **abstract** that gives an executive overview over the document.
3. An **introduction** that leads the reader into the topic, discusses the problems solved in the document and their relation to the "real world", and generally argues that reading the paper is worth the reader's attention.

4. A **preview**, which outlines the structure of, the contributions in, and methods used in the document.
5. A discussion of the **state of the art** on the topic of document.
6. The establishment of a **common ground** between the reader and the author, which *i*) recapitulates or surveys concepts and results from the documents/knowledge commons to make the document self-contained (for its intended audience) *ii*) identifies any assumptions and gives the ensuing contributions a sound terminological basis.
7. The **contributions** part, which contains the development of new knowledge in form of e.g. new insights, new interpretations of known concepts, new theorems, new proofs, new applications/examples or new techniques of achieving results.
8. An **evaluation** of the contributions in terms of applicability or usability.
9. A discussion of **related work** which reviews the contributions and their relation to existing approaches and results from the literature.
10. A **conclusion** which summarizes the contribution with the benefit of hindsight and relates it to the claims made in the introduction.
11. Literature references, an index, a glossary, etc. and possibly appendices that contain material deemed supplementary to the contributions.

Even though the form or order of the structural elements may vary over publication venues, and certain elements may be implicit or even missing altogether, the overall structure is generally stable.

It may be surprising that only one in eleven parts of a mathematical document – the "contributions" – arguably the largest – is fully dedicated to transporting the payload of the paper. All other contribute to either the dissemination[1], understanding[2] and aggregation processes. We will see that the latter is mainly driven by the common ground (point 6. above), which we will analyze in more detail next.

### 2.2 Common Ground/Recapitulation in Mathematical Research

To get an overview over recaps in the literature, we randomly selected 30 papers from the new submissions to `http://arxiv.org/archive/math` and analyzed their structure. All had a significant common ground section that recapitulates the central notions and fixes notations. We show two examples where the mathematics involved is relatively elementary.

*Example 1.* [HK15] discusses covers of the multiplicative group of an algebraically closed field which are formally introduced in the beginning of the paper as follows:

[1] 1. for referencing, 2. for determining interest

[2] 3. and 10. for broader context, 5. and 9. for problem context, 4. for document navigation, 8. for assessment of value, and 11. for further reading

> **Definition 1.1** Let $V$ be a vector space over $Q$ and let $F$ be an algebraically closed field of characteristic $0$. A *cover of the multiplicative group* of $F$ is a structure represented by an exact sequence $0 \to K \to V \to F \to 1$ , where the map $V \to F^*$ is a surjective group homomorphism from $(V, +)$ onto $(F, \cdot)$ with kernel $K$. We will call this map *exp*. (1)

However, later, the authors source the concept origin to an earlier paper ("[13]") and effectively import the terminology, definitions and theorems. For instance, when establishing results, [HK15] mentions "Moreover, with an additional axiom (in $L_{\omega_1\omega}$) stating $K \cong Z$, the class is categorical in uncountable cardinalities. This was originally proved in [13] but an error was later found in the proof and corrected in [2]. Throughout this article, we will make the assumption $K \cong Z$.".

In the second example, the situation is a bit more complex, since the import of the terminology and definitions is not direct, but involves a choice.

*Example 2.* [Bar15] studies the properties of multinets. In the preliminaries section they are introduced with the following definition:

> **Definition 2.1** The union of all completely reducible fibers (with a fixed partition into fibers, also called blocks) of a Ceva pencil of degree $d$ is called a $(k, d) - multinet$ where $k$ is the number of the blocks. The base $X$ of the pencil is determined by the multinet structure and called the base of the multinet. (2)

Later in that section some properties of multinets are introduced with the phrase "Several important properties of multinets are listed below which have been collected from [4,10,12].". The referenced papers all use slightly different definitions of multinets but they are assumed to be equivalent so that the properties hold. In fact, in this paper ([Bar15]) the assumption is made explicit – although not proved – from the start: "There are several equivalent ways to define multinets. Here we present them using pencils of plane curves."

The next example is not from our 30 examples, since we want to show an even more complex situation.

*Example 3.* [CS09] studies the halting problem for accelerated Turing machines and starts off the discussion with an informal introduction of the topic.

> An accelerated Turing machine (sometimes called Zeno machine) is a Turing machine that takes $2^{-n}$ units of time (say seconds) to perform its $n^{\text{th}}$ step; we assume that steps are in some sense identical except for the time taken for their execution. (3)

This is a telegraphic version of the full definition, which is given in the literature. Actually [CS09] continues with an overview of the literature, citing no less than 12 papers, which address the topic of accelerated Turing machines. One of these supposedly contains the formal definition, which involves generalizing Turing machines to timed ones, introducing computational time structures, and singling out accelerating ones, e.g. using (4).

**Definition 1.3**: An **accelerated Turing machine** is a Turing machine $M = \langle X, \Gamma, S, s_o, \square, \delta \rangle$ working with with a computational time structure $T = \langle \{t_i\}_i, <, + \rangle$ with $T \subseteq \mathbb{Q}_+$ ($\mathbb{Q}_+$ is the set of non-negative rationals) such that $\sum_{i \in \mathbb{N}} t_i < \infty$. (4)

Note that the definition of an ATM [CS09] is an instance of definition 1.3, which allows arbitrary time structures.

### 2.3 Secondary Literature: Education/Survey

A similar effect can be observed with educational materials or survey articles, whose concern is not to make an original contribution to the knowledge commons, but to prepare a document that helps an individual or group study or better understand a body of already established knowledge. Consider for instance, slides and background materials (lecture notes, text books, encyclopaedias), where the slides often have telegraphic versions of the real statements, which verbalize more rigorous definition.

This is illustrated in Example 4 which is inspired from the notes of a first year computer science course taught by the first author. The example is a simplified and self-contained version of the original which in itself is only one instance of a commonly occurring pattern in the course notes.

*Example 4 (A Course grounded in a Formal Library).* A course which introduces (naive) set theory informally, but grounds itself in a formal, modular definition. In the cited source, we have a careful introduction in the form of a modular theory graph starting at a theory that introduces membership relation and the axioms of existence, extensionality, and separation and defines the set constructor $\{\cdot|\cdot\}$ from these axioms. In the course notes we have a theory that "adopts" the symbols $\in$ and $\{\cdot|\cdot\}$ but not the associated axioms. Instead it "defines" them by alluding to the intuitions of the students. Then the course notes continue with introducing set operations ranging from set union to the power set.

We observe that course notes in Example 4 are self-contained in the sense that they can be understood without knowing about the formal development. This self-containedness is important intra-course didactics. But it also has the problem that the courses become insular; how are students going to communicate with mathematicians who have learned their maths from other courses? This is where alluding to the literature comes in, by connecting the course notes with it.

*Example 5.* The situation in mathematical textbooks is similar in structure to that in research papers –perhaps more pronounced. Consider the following passage from Rudin's classical introductory textbook to Functional Analysis [Rud73, p. 6f].

**1.5 Topological spaces** A *topological space* is a set $S$ in which a collection $\tau$ of subsets (called *open sets*) has been specified, with the following properties: $S$ is open, $\emptyset$ is open, [...] Such a collection is called a *topology* on $S$. [...] (5)

This is continued later – vector spaces have been recapped earlier in section 1.4 – with:

> **1.6 Topological vector spaces** Suppose $\tau$ is a topology on a vector space $X$ such that
> (a) *every point of $X$ is a closed set, and*
> (b) *the vector space operations are continuous with respect to $\tau$*
> Under these conditions, $\tau$ is said to be a *vector topology* on $X$, and $X$ is a *topological vector space*. (6)

Note that Rudin does not directly cite the literature in these quotes, but in the preface he mentions the vast literature on function analysis and in Appendix B he cites the original literature for each chapter. The situation in textbooks is also different from research articles in that textbooks – like survey articles, and by their very nature – do not add new knowledge or new results, but aggregate and organize the already published ones, possibly reformulating them for a more uniform exposition. But still, one can distinguish recap parts – as the ones above – which are much more telegraphic in nature from the primary material presented in the textbook.

### 2.4 Common Ground in Formal Mathematics

Where applicable, common ground in formal mathematics is typically established via direct imports of symbols, theorems, notations, etc. Formal documents emphasize correctness and do not focus on human readability so they do not reintroduce concepts or provide, verbalizations of definitions.

For instance, In Isabelle and Coq knowledge is organized in *Theories* and *Modules* which are effectively named sets of declarations. The incremental development process is enabled via the IMPORTS and, respectively, REQUIRE IMPORT statements that effectively opens a library module by name and enables its declarations to be used in the current development.

In Mizar, formal documents (called *articles*) can be exported as PDF files in a human readable format. The narrative documents contain a part that verbalizes the imports from the source documents and the notation reservations which can be seen as a common ground section.

*Example 6.* The common ground part for [RK13b]

> The notation and terminology used in this paper have been introduced in the following papers: [4], [11], [12], [19], [9], [3], [5], [6], [21], [22], [1], [2], [7], [18], [20], [24], [25], [23], [16], [13], [14], [10], [15], and [8]. [...] In this paper $T$ , $U$ are non empty topological spaces, $t$ is a point of $T$ , and $n$ is a natural number. (7)

## 3 Publication and Dissemination in Theory Graphs

In this section we look more closely at the examples from Section 2 and how each can be represented using theory graphs. But first, we look at the aspects

common to all examples to form an intuition of the theory graphs structures that are needed.

The examples in Section 2 are each slightly different but they have fundamental common aspects. First, each paper starts with establishing a common ground on which the results of the paper are built. This leverages the literature in two ways.

- Firstly, concepts from the literature are used to conveniently build up the local definitions. From the theory graphs perspective this functions as a (possibly partial) import.
- Secondly, properties of locally introduced concepts are *adopted* from the literature. Mathematically, this is justified by and (implicit or explicit) subsumption between the local definition and that used by the referenced theorem. From the theory graph perspective this function as a theory morphism that induces the properties locally due to its truth-preserving semantics.

Therefore, a paper corresponds, not to a single theory, but to a theory pattern that leads to a theory of the main contribution of the paper.

Secondly, the notion of "literature" and the existence of concepts beyond a particular definition (so that equivalent definitions imply one is talking about the same platonic concept) are common to all examples. We believe that what happens in mathematical practice is that definition and foundational choices are abstracted away as implementation details and the important concepts and their properties are used as an interface to each theory (in the mathematical sense, e.g. group theory). But this is precisely the situation that realms try to capture in theory graphs. Therefore, we maintain that, from a theory graph perspective, informal mathematical papers refer (and contribute to) realms rather than individual theories.

### 3.1 Realms

Intuitively, a realm [CFK14] is a theory structure in a theory graph $G$ (i.e. a subgraph of $G$) that abstracts from the development and provides practitioners with the useful symbols and theorems via an *interface theory*.

We briefly introduce realms and the background concepts below and refer to [CFK14] for details.

First, in the following, *theories* are named sets of declarations (i.e. symbols, axioms or theorems). Additionally, *theory morphisms* (or *views*) are truth-preserving mappings from a source theory to a target theory and formalize inheritance and applicability of theorems. Theories can access and use declarations from other theories by importing them, either directly (*plain includes*), or via a translation (*structures*).

An important concept for realms is that of a *conservative extension* which usually occurs when a theory includes another and contains only theorems and derived symbols (i.e. adds no axioms or primitive symbols). An essential property of conservative extensions is that if $S'$ is a conservative extension of $S$ then there

is view $v$ between $T$ and $S$ iff there is a view between $T$ and $S'$ in the same direction. In fact, we will often talk about views *modulo conservativity* below.

Figure 1 shows a prototypical realm with $F$ as its interface theory (also called a *face*) and $n$ pillars each representing a different (yet equivalent) development of the concepts in the face. Common examples are the different ways to define natural or real numbers. Each pillar is a conservative development in the sense that all theories in a pillar are conservative extensions of a bottom theory (denoted with $\bot$). A top theory (denoted with $\top$) aggregates all symbols, axioms and theorems declared within the pillar. The view pairs at the bottom establish the equivalence of the pillars and the $n$ views $I_k$ capture the relation of interface-implementation between the face and each pillar.

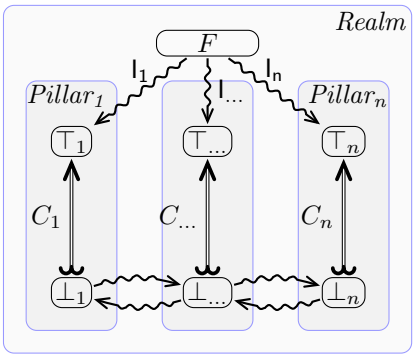


Fig. 1: The Architecture of a Realm

### 3.2 Realms as a Model for Dissemination & Aggregation

Figure 2 shows the general case for the representation of a paper as part of a theory graph. The "literature" for the mathematical theory to which the paper contributes is represented as a realm with a face and several pillars. The paper references a document within the field, that is naturally part of a pillar and grounds the recap theory. The contribution of the paper is a theory in itself that includes the recap theory and is a conservative extension of it. Again, the fact that we are representing the contribution in a single theory is a simplification for presentational simplicity which does not lead to a loss of generality. The view $v$ ensures that the paper can make use of concepts and theorems from the realm, as they can be accessed via $v$.

In our analysis we first restrict ourselves to the case where there is a single recap for simplicity and expositional clarity. This already covers the majority of research papers we have analyzed; they mainly build on one earlier paper and extend it. Indeed, all three examples from Section 2.2 fall into this category, they import the definitions and terminology from a central cited paper, but call on others from the same realm for results, context, and support.

We recognize four special cases for (single) recaps based on the nature of $r$ and discuss each individually below. First we have to decide the home theory of the symbols that the recap introduces. If the home is the cited theory then $r$ is an import and we have a *plain recap* (3.3). Otherwise, we have new symbols in the recap theory that are somehow related with the ones in the cited one. In that situation we have three sub-cases depending on the relation between the recap and cited theory: *equivalence recap* (3.4), *specialization recap* (3.5) and, in the informal case, *postulated recap* (3.6).

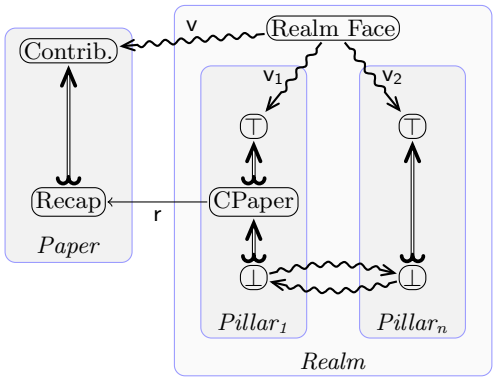



Fig. 2: General Case for Recaps

Finally, we have the case where the paper builds on several others and has *multiple recaps* (3.7).

### 3.3 Special case: Plain Recaps

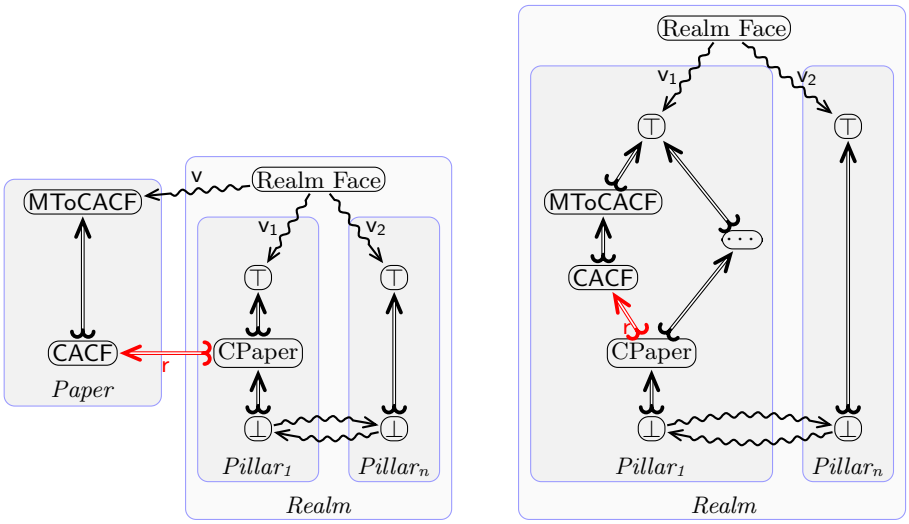
One situation is that of plain recaps where the relation $r$ is an inclusion into the recap from the cited paper. Typically the include $r$ is a conservative extension of the cited paper. For instance the "covers of the multiplicative group" from Example 1 directly uses the concept from the cited paper (CPaper), but gives a concise verbalization of its definition. This allows it to make use of the results in two other papers higher up in the pillar of the cited paper. The situation is shown in Figure 3a. Note that, if $r$ is conservative, then we have a **pillar extension** for the realm which justifies the new paper becoming part of the realm's literature (see Figure 3b). It also makes $v$ exist as induced by $v_1$ modulo conservativity.

Plain recaps can also model the formal examples (e.g. Example 6) but in that situation it is not too interesting as we have the degenerate case for the realm itself.

### 3.4 Special Case: Equivalence Recap

Another common situation is that of equivalence recaps where the relation $r$ is an equivalence (isomorphism) between the two theories. We can represent the relation $r$, in this case, as two views $v_{to}$ and $v_{from}$, one in each direction between the recap and the cited paper that ensure their isomorphism. Then, the view $v$ is induced by $v_{from} \circ v_1$ modulo conservativity. Moreover, the contribution of the paper carries over to the realm via the view $v_{to}$.

This occurs, for instance, in Example 2 where this intuition is explicitly written down in the paper as "There are several equivalent ways to define multinets." (although not proved). In fact it is the most common situation in the sample papers we studied.




(a) Publication Graph (b) Aggregation Graph

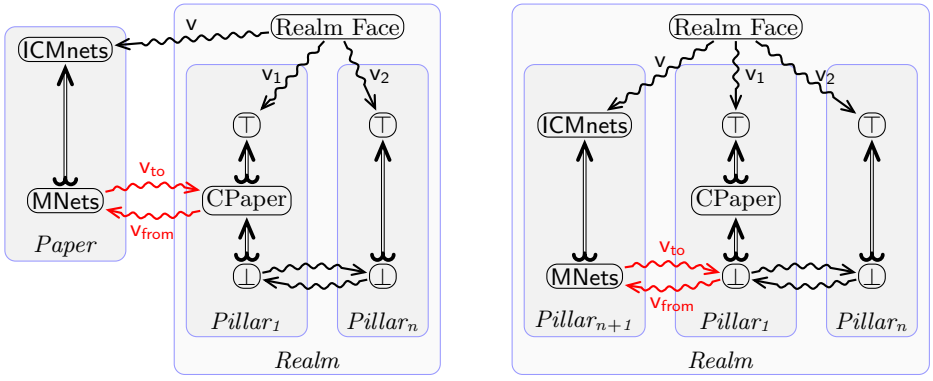
Fig. 3: Plain Recaps (Example 1)

Note that adding an equivalent definition corresponds to a **realm extension**, where the face is fixed, and the view from the face to the current theory can be postulated. Therefore, in Figure 4a the paper effectively extends the realm (or the current pillar) as introduced in Section 3.1. This corresponds to the mathematical practice of "contributing to" a field (or mathematical theory). This resulting realm after knowledge aggregation is shown in Figure 4b, where the new paper contributes a new pillar to the realm. The equivalence is ensured by $v_{from}$ and $v_{to}$ as we take into account conservativity to reduce them to the $\bot$ theory.

### 3.5 Special Case: Specialization Recap

Thirdly, we have the case where $r$ is a specialization relation that can be represented as a view $v_{from}$ from the cited theory to the recap. Same as in the previous case, this ensures the existence of $v$ as $v_{from} \circ v_1$ modulo conservativity. However it does not directly contribute the results of the paper back to the (same) realm as they concern only a special case of the concepts in the realm.

This is the case in Example 3 where the definition from the paper is a specialization of the one in the literature. In [CS09], the definition of the accelerated Turing machine involves a concrete step size ($2^{-n}$), whereas the definition it recaps allows arbitrary sequences of step sizes as long as their sum remains finite. Thus we have the situation in Figure 5. Theory ATM contains the (opaque) sentence (3), but there cannot be a view from ATM to atm as that is more general.




(a) Publication Graph (b) Aggregation Graph

Fig. 4: Equivalence Recaps (Example 2)

But we do have a view to $\mathsf{atm}(2^{-n})$, which naturally arises in treatments of accelerated Turing machines as an example. That special case can form a realm of its own, namely the realm of accelerated Turing machines with step size $2^n$. Then we can talk about aggregation with that realm (via the view $\mathsf{v_{to}}$) but we omit that here for simplicity – the aggregation is similar as for equivalence recaps, except with the specialization realm.

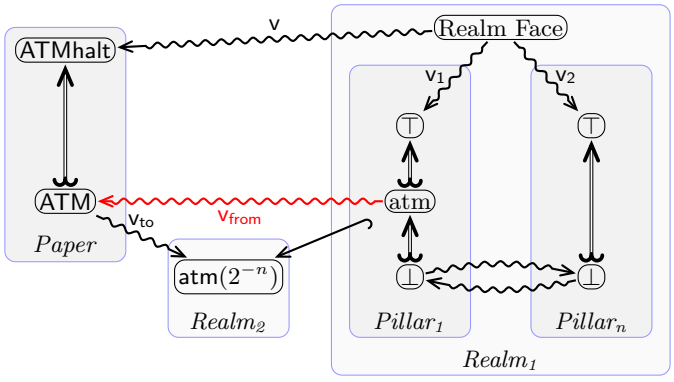



Fig. 5: Publication Graph for Specialization Recaps (Example 3)

### 3.6 Postulated Recap/Adoption

Finally, we have the case for educational material such as the one in Example 4 where $r$ cannot be directly modeled as either an include or a view. This is caused

by the constraint of self-containedness of such materials. Normally, in the case where a more formal development is used we could represent it as an include and be in the case for plain recaps. However, the home theory of the new symbols must be the current development in order for it to be self-contained, so we cannot use an include. Instead we envision a special kind of import that *adopts* the included symbols effectively changing their home theory to the current one. But, then the view $v$ is not justified so we must also assert its existence. In that case we call $v$ a *postulated* view and the the relation $r$ is an *adoption* (see Figure 6). We leave working out the precise details of postulated views and adoptions in flexiformal theory graphs for future work.

This is the situation in Example 4 where the recap theory SET includes only the symbols $\in$ and $\{\cdot,\cdot\}$ from the formal development ZFset, but not their axioms. Instead the symbols are "defined" by alluding to the literature (common knowledge). We claim this verbalization effectively postulates the existence of $v$, by implying that the semantics of the two symbols is compatible with that given in the literature (which we represent as a realm).

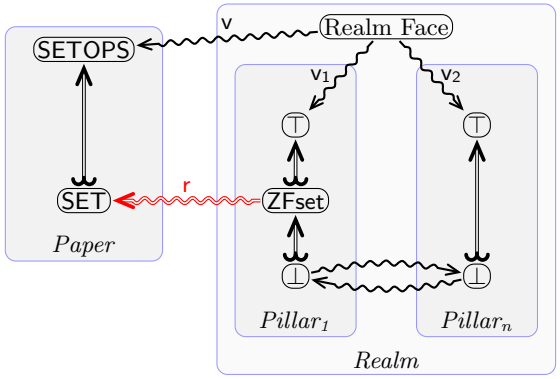


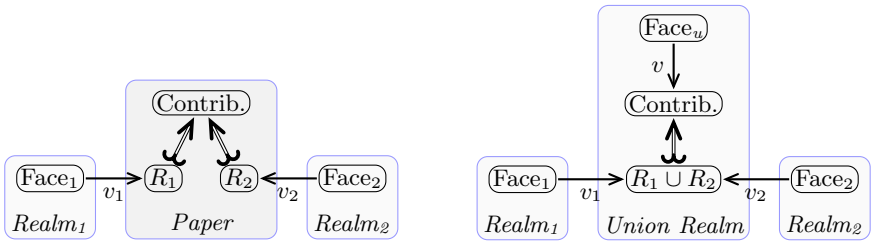

Fig. 6: Publication Graph for Generalization/Unspecified Recaps (Example 4)

Note that we omit the aggregation part for this case as the purpose of such educational or survey material is typically to provide a concise overview of a realm rather than to contribute to it.

### 3.7 Multiple Recaps

Up to now we have only treated cases with single recaps to ease the exposition. But papers and especially textbooks often recap from different realms and base the rest of the exposition on them.

This is a the situation on the left of Figure 7; for the aggregation phase this begs the question where the contribution should be placed. In the recap in Rudin's book mentioned in Section 2.3 we have separate recaps of vector spaces




(a) Publication Graph (b) Aggregation Graph

Fig. 7: Multiple Recaps (Example 5)

and topological spaces (5), and we analyze them as theory morphisms from their respective realms. In this case, there is the realm of topological vector spaces (6) which imports from both realms, this is the natural place for the contributions. In the case such a realm does not exist yet, the paper can be used as the natural starting point for (first pillar of) the realm. Actually, the "union realm" concept in Figure 7 is a bit simplified. The contribution of the paper will usually add some conditions – like conditions (a) and (b) in (6) – and use that for the base theories of the realms. This does not invalidate our claim that there is always a natural realm – which may have to be created – for the contribution of the "paper".

## 4 Conclusion and Future Work

We have presented a flexiformal model of the mechanics of paper-based dissemination of research results and their aggregation into a structured knowledge commons. We model the latter as an underlying theory graph structured by inclusions and views that is further structured into a graph of realms to abstract from details of the particular low-level developments of the mathematical domains.

We identify the recap+contribution structure in mathematical papers as the mechanism by which papers can at the same time be made self-contained for human readers and by which the contribution can be integrated into the knowledge commons: the recap anchors the contribution in the commons. It is the realms structure with its equivalent pillars and abstraction capabilities that gives the recaps the necessary flexibility to adequately model the variety of anchors we see in mathematical documents.

We have validated our model by identifying the recaps and their types in 30 recent papers randomly selected from a preprint archive. To obtain a more scientific evaluation of the model, we need a much larger and more varied sample. We are currently developing an annotation ontology for realms and recaps for the KAT annotator [Dum+14] as a basis for a more principled and sustainable analysis. This will also give us the data to develop our model further.

In the future we want to look into the communication-enabling partial isomorphisms postulated in Section 3.6 and see whether [KRSC11] is directly applicable.

We believe that the realms-based model can be extended to handle recaps from multiple realms in one document. For the document model, this is not a problem, since we would just have multiple bases for the conservative development. For the aggregation things become more complex. Intuitively, the contribution must be integrated into a realm that is the "union" of the realms, and if that does not exist yet, the realm can be initialized with the paper at hand.

An implementation of realms in the MMT API [Rab13] is under way, this will allow us to validate the model proposed in this paper from the synthetic direction: If we have a realm-structured knowledge commons, then we may be able to auto-generate recaps and common ground sections to obtain narrative presentations of fragments that are more self-contained and readable to the human reader. This is particularly interesting for the concept of "guided tours" in content-based eLearning systems: auto-generated explanatory narratives leading to a given mathematical concept by topologically sorting the dependency relation given by the theory graph in the content commons. For the "early" parts on the border to the estimated common ground, recaps might be more suitable than direct copies of the definitions.

*Acknowledgements* This work has been supported by the Leibniz Association under grant SAW-2012-FIZ_KA-2 and the German Research Foundation (DFG) under grant KO 2428/13-1.

## References

[Bar15] Jeremiah Bartz. "Induced and Complete Multinets". In: *ArXiv e-prints* (Feb. 2015). arXiv: 1502.02059 [math.AG].

[CFK14] Jacques Carette, William Farmer, and Michael Kohlhase. "Realms: A Structure for Consolidating Knowledge about Mathematical Theories". In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. MKM Best-Paper-Award. Springer, 2014, pp. 252–266. URL: http://kwarc.info/kohlhase/submit/cicm14-realms.pdf.

[CS09] Cris Calude and Ludwig Staiger. *A Note on Accelerated Turing Machines*. CDMTCS Research Report 350. Centre for Discrete Mathematics and Theoretical Computer Science, Auckland University, 2009. URL: http://www.cs.auckland.ac.nz/CDMTCS/researchreports/350cris.pdf.

[Dum+14] Mircea Alex Dumitru et al. "System Description: KAT an Annotation Tool for STEM Documents". 2014. URL: http://kwarc.info/kohlhase/submit/cicm14-kat.pdf.

[HK15] Tapani Hyttinen and Kaisa Kangas. *On model theory of covers of algebraically closed fields*. 2015. URL: http://arxiv.org/pdf/1502.01042.pdf (visited on 02/16/2015).

[JFM] *Journal of Formalized Mathematics*. URL: http://www.mizar.org/JFM (visited on 09/27/2012).

[Koh06] Michael Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[Koh13] Michael Kohlhase. "The Flexiformalist Manifesto". In: *14th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012)*. Ed. by Andrei Voronkov et al. Timisoara, Romania: IEEE Press, 2013, pp. 30–36. URL: http://kwarc.info/kohlhase/papers/synasc13.pdf.

[KRSC11] Michael Kohlhase, Florian Rabe, and Claudio Sacerdoti Coen. "A Foundational View on Integration Problems". In: *Intelligent Computer Mathematics*. Ed. by James Davenport et al. LNAI 6824. Springer Verlag, 2011, pp. 107–122. URL: http://kwarc.info/kohlhase/papers/cicm11-integration.pdf.

[MizLib] *Mizar Mathematical Library*. URL: http://www.mizar.org/library (visited on 09/27/2012).

[Rab13] Florian Rabe. "The MMT API: A Generic MKM System". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics. (Bath, UK, July 8–12, 2013). Ed. by Jacques Carette et al. Lecture Notes in Computer Science 7961. Springer, 2013, pp. 339–343. DOI: 10.1007/978-3-642-39320-4.

[RK13a] Florian Rabe and Michael Kohlhase. "A Scalable Module System". In: *Information & Computation* 0.230 (2013), pp. 1–54. URL: http://kwarc.info/frabe/Research/mmt.pdf.

[RK13b] Marco Riccardi and Artur Kornilowicz. "Fundamental Group of $n$-sphere for $n \geq 2$". In: *Formalized Mathematics* 20.2 (2013), 97–104. DOI: 10.2478/v10037-012-0013-1.

[Rud73] Walter Rudin. *Functional Analysis*. McGraw Hill, 1973.

[Wen07] Makarius Wenzel. "Isabelle/Isar — a generic framework for human-readable proof documents." In: *From Insight to Proof: Festschrift in Honour of Andrzej Trybulec*. Ed. by R. Matuszewski and A. Zalewska. Vol. 10:23. Studies in Logic, Grammar and Rhetoric. University of Białystok, 2007, pp. 277–298. URL: http://mizar.org/trybulec65/.