Creating a Minimal Information Vocabulary for a Reproducible Method Description
A Case in Column Chromatography

Dena Tahvildari
Computer Science Department
VU Amsterdam
The Netherlands
Email: d.tahvildari@vu.nl

Anne Vissers
Laboratory of Food Chemistry
Wageningen University
The Netherlands
Email: anne.vissers@wur.nl

Guus Schreiber
Computer Science Department
VU Amsterdam
The Netherlands
Email : guus.schreiber@vu.nl

Jan Top
Food & Biobased Research
Wageningen UR
and VU Amsterdam
The Netherlands
Email : jan.top@wur.nl

***Abstract*—Descriptions of experimental methods in scientific publications are often incomplete or inadequate. In these cases, the experimental work cannot be reproduced or verified due to lack of information. To facilitate the documentation of lab methods, in some domains minimum information guidelines have been developed. If implemented, these guidelines ensure that the information about the method can be easily verified, analysed and clearly interpreted by a wider scientific community. However, there is an evident lack of automated documentation tools to create and edit laboratory reports that follow these guidelines and at the same time do not impose a too rigid framework on the scientist. This paper describes the very first step towards the development of semantically rich but free-text editor for creating descriptions of experimental methods. We created and evaluated the vocabulary for reporting a column chromatography experiment, which is developed using the MIAPE guidelines. Our goal is to check if we can use the MIAPE guidelines in the food chemistry domain.The ultimate use of the vocabulary is in semantically enriched editorial software. An editor should give knowledge-based guidance to the author and semi-automatically add meta-data. The first step in designing such editor is to construct supporting vocabularies and evaluate their use in the domain of interest. Our initial application domain is laboratory of food chemistry.**

Keywords–MIAPE; vocabulary; material and method sections; HPLC; reproducibility; laboratory experiments.

I. Introduction and objective

Transparency and reproducibility are recognized as essential features of science [1][2]. The quality of methodology descriptions are important factors for transparency and reproducibility. Therefore, providing adequate research documentation is an important task of a scientist.

In this paper, we discuss the very first step towards creating a semantic support for writing a reproducible method description, which intends to allow researchers to perform this task effectively and efficiently. The notion of research reproducibility has different interpretations, varying between different research fields. Research reproducibility commonly implies that, as an ultimate product of scientific investigation, research papers must be accompanied by a detailed description of the computational or experimental environment that are used to produce the result. According to Clarebout's principle [3] "An article [...] in a as a means for scholarly communication is not the scholarship itself; it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions that generates the results". This idea promotes that research data, algorithms, codes and protocols are not simply ancillary information, but first class scholarly products as important as the paper itself. We define the term reproducibility as "the ability to investigate a phenomenon using the similar conditions as in the original experiment". We emphasize that the conditions do not need to be identical, but only similar, since slight variations are essential for scientific understanding of a phenomenon.

We focus on reproducibility in the context of laboratory research. There can be two reasons for an experiment not to reproduce the same phenomenon:

1) the hypothesized mechanism does not manifest itself, even having all conditions right (falsification)
2) the conditions under which the hypothesized effect can manifest itself have not been adequately fulfilled

The second condition can result from a poor description of the experimental conditions. This is why the scientific method requires explicit records of "all" experimental conditions. Having a report of the precise experimental process, and data is necessary to explain why some result has been found, why results could be different or be same as the results found in a different condition. For repeating a mechanism, it is important to know which assumptions and conditions must hold for the mechanism to manifest itself. In addition to serving scientific integrity, another reason for having details of an experiment is to make the transition to applications. For example, a Standard Operating Procedure (SOP), is intended as a step-by-step instruction to achieve a predictable, standardized, desired result, often within the context of a longer overall process.

Although a full account of the experimental conditions would be ideal, this cannot be achieved in practice. It is not possible to describe literally all details that possibly might be of influence; what's worse, scientists are usually not inclined to allocate time and effort to the "administrative" task of creating extensive documentation. Transparency in documentation is costly for scientists – in terms of time and effort. Taking into account that only the "essential" conditions need to be registered, the question is how researchers can be supported to realize which are these essential conditions that are sufficient to perform a "similar" experiment. When asked for, most scientists embrace transparency and reproducibility as disciplinary norms and values of science [4]. Therefore, one might expect that providing full documentation of methods and data is routine in daily practice. Yet, a growing body of evidence suggests that this is not the case [5][6]. It is becoming increasingly clear that the current publication model falls short in promoting transparency and reproducibility. A recent Nature report from researchers at the Amgen corporation showed that only 11% of the academic research in the literature are reproducible. Individual motivation and personal efficiency gain are other variables in promoting reproducibility and transparency of scientific methods [7][8].

In the present publication model, the conditions under which an experiment is performed are described in the "material and method" section of a scientific article. This section should provide information about the materials, procedures and critical steps that are used in the course of an experiment, such that the procedure can potentially be reproduced as faithfully as possible [9].

Minimum information guidelines (MIAPE) have been developed in various research domains to facilitate documentation [10]. Although they provide valuable guidance for reporting the necessary information about the method, they are rather high level, and do not give the detailed and context-specific support that is needed at the time of writing a method description. We think that a semi-automated use of minimum information guidelines in editorial applications could improve the quality of laboratory reports and method sections of publications, while limiting the time and effort needed to produce these.

Our approach to solve the quality problem of laboratory method reports and (potentially) lab protocols – in terms of transparency and reproducibility – relies on the use of the Semantic Web technologies and formal methods. We believe that in order to provide support for scientific authors, the first step would be to create a formal model of the underlying domain knowledge. A structured vocabulary or ontology can help to provide context-dependent suggestions to authors. We emphasize that we do not address the quality of the argumentation followed, nor the soundness of the research method.

In this study, we start off by exploring the minimum information guidelines for reporting a column chromatography technique in the food chemistry domain. Our hypothesis is that terms occurring in the guidelines should be present in the method sections of published papers. In the second section of this paper, we present relevant literature regarding the problem, and current approaches. In the third section of the paper, we briefly familiarise our readers with high performance liquid chromatography techniques as the first case study. Our approach to create the first draft of the vocabulary is presented in section 4. Section 5 is dedicated to results of the term frequency measurements. Finally, we discuss the results and provide some hypotheses for further testing in section 6.

II. Related work

Several initiatives have identified the problem of inadequate reporting and have proposed solutions. The National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3R) assessed methodological reports in the literature for in-vivo research. They evaluated 271 publications and showed that only 60% of the articles included information about the number and characteristics of the animals (strain, sex, age, weight) and approximately 30% of the articles lacked detailed descriptions of the statistical analyses used. Built upon this study, the ARRIVE [11] [12] guidelines were developed for reporting in-vivo experiments, pertaining to animal research.

To promote scientific reproducibility, the FORCE11 community has published a set of recommendations for minimal data standards for biomedical research and published a manifesto to improve research communication. The BioSharing initiative contains a large registry of community standards for structuring and curating data sets. It has made significant strides towards the standardization of data via its multiple partnerships with journals and other organizations [13].

The most relevant work to our research is an initiative in the Proteomic community. The problem of accurate methodological reporting is addressed by developing the minimum information documentation guidelines (MIAPE guidelines) as a standard, along with the development of MIAPE-supported software tools. For example, the ProteRed MIAPE Web toolkit was developed to fulfill the lack of bio-informatics tools to create and edit standard file formats and reports. It allows these to be embedded in proteomics research work flows. This system is able to verify if the report fulfills the minimum information requirements of the corresponding MIAPE modules while highlighting missing information and inconsistencies in a report. In other words, this system works as a MIAPE compliance checker and has been designed to support the validation of experimental meta-data [14].

Our approach is similar to the ProteRed compliance checker in terms of using semantics. However, we intend to develop MIAPE-CC vocabularies and use it in editorial applications that are frequently used by scientists, such as Microsoft Word. We believe that in order to enable researchers to provide a reproducible method description with low cost, we need to develop a knowledge base of reporting requirements and apply them in the most frequently used scholarly communication tools [15].

III. Case description

This section describes high-performance liquid chromatography (HPLC). We have selected this technique as a use case to build a vocabulary and select reference articles. HPLC is a chromatographic method that is used to separate a mixture of compounds in analytical chemistry and biochemistry so as to identify, quantify or purify the individual components of the mixture. HPLC can be used in the following applications, on small scale (analytical) and large scale (preparative):

1) Mixture characterization (analytical)
2) Water purification (preparative)
3) Pre-concentration of trace components (preparative)

Examples of HPLC chromatography types are:

1) Ion-exchange chromatography of proteins
2) Ligand-exchange chromatography
3) Reversed phase chromatography
4) Size exclusion chromatography

The sample mixture to be separated and tested is sent into a stream in the mobile phase percolating through the column. There are different types of columns available with sorbents of varying particle sizes and materials. For most types of chromatography, the mixture has interaction with the sorbent, also known as the stationary phase. The separation depends on the balance between compound affinities for the sorbent (As) and for the mobile phase (Amp). To separate compounds, a constant flow of mobile phase over the column is applied, which changes in composition gradually. When “As” is less than “Amp”, the compound detaches from the sorbent and travels in the mobile phase stream towards the detector. The time that the compound needs to emerge at the detector is referred to as the retention time. For each component in the mixture, this depends on its chemical nature, the characteristics of the column and the composition of the mobile phase. Changes in these conditions yield different retention times. The retention time is measured under specific conditions and together with data from specific detectors used, is considered as the identifying characteristic of a given analyte.

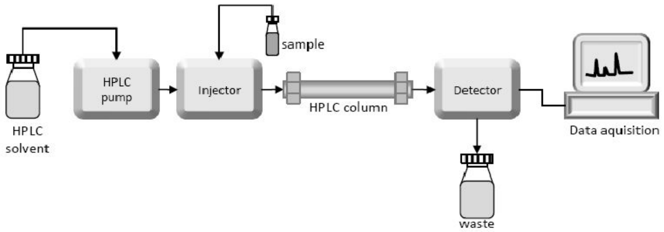

Figure 1. Components of a chromatographic process

The preparation of the mobile phase affects the quality of separation. The mobile phase might contain acids like formic, phosphoric or trifluoroacetic acid or salts to force components into their non-charged states and increase column retention. A pump is used to generate a specified flow of the mobile phase. Although manual injection of samples is still possible, most HPLC systems are now fully automated and controlled by computer software (e.g., XCalibur). The injector, or auto sampler, introduces the solvent into a phase stream that carries the sample into the column, which is under high pressure and contains specific packing material needed to affect separation. The packing material is referred to as the stationary phase because it is held in place by the column hardware. A detector is used in these experiments to see the separated compound bands width as they elute from the column. The information is sent from the detector to a computer software which generates the chromatogram. The mobile phase exits the detector and is either discarded as waste in analytical chromatography, or collected in case of preparative chromatography. Figure 1 schematically presents a the fundamental components of a chromatographic process.

In an HPLC experiment, data about cleaning the column, system calibration, retention time, sample components, and graphs that are generated by the HPLC system and the detector are analysed and documented in XCalibur. This software enables scientists to gather, analyse, visualize the information about the chromatogram. Although XCalibur has features for creating metadata about the experiment, we have observed that it’s added value for documenting the experiments, has not been fully realized by the scientists. The reason is not known to us; however, we are interested to understand the functionality and usability of this software for the future use.

IV. Method

This section describes how we extracted terms from the guidelines and how the resulting vocabulary was evaluated in the food chemistry domain. We should indicate that the “quality” of the vocabulary is determined by the degree to which it assists scientists in the considered domain to create reproducible method descriptions in an efficient manner. However, the first measure for the quality is the extent to which the terms contained in the vocabulary occur and convey the intended meaning in published method descriptions. We explicitly do not use the “method sections” as sources for the creation of vocabulary, but only for the evaluation of our vocabulary in the domain of Food Chemistry. This is to guarantee the independence of our method from the specific set of method sections we selected. The MIAPE-CC is our starting point for creating a vocabulary in the considered domain [16]. Table I presents the seven categories, each representing an essential part of an experimental setup. It covers a column chromatography experiment from the selection and configuration of a column, through the selection of a suitable mobile phase and verification of the relevant performance characteristics, to the collection of fractions and associated detector readings. We manually extracted the main concepts from the guideline. In total, 83 terms were extracted. Table II provides an example of the main categories and the associated terms. In the next step we measured the occurrence of the extracted terms in the material and method sections. The library of Wageningen University (The Netherlands) kindly provided us a list of articles from laboratory of food chemistry that cover five predefined criteria.

TABLE I. The MIAPE-CC CLASSIFICATION

MIAPE-CC CLASSIFICATION	
Class	Description
global descriptors	All the general information about the experiment, such as the date on which the work described was initiated and etc
sample	Description of the source, such as means of collection, volume, concentration or previous step of processing.
equipment	Description about the type of column and the chromatography system that are being used.
mobile phase	The mobile phase is the phase that moves in a definite direction. It may be a liquid, or a gas (GC), or a super critical fluid.
column run process	The total time of the column run with appropriate units.
pre and post run process	a description of the purpose of the process, such as equilibration, calibration or washing (this may be part of the column run, as one step or as preconditioning of the column prior to use).
column output	a description about the output that is selected for detection and/or fraction.

TABLE II. EXAMPLE TERMS EXTRACTED FROM MIAPE-CC

general description	equipment	sample	mobile phase
date stamp	column	name	name
responsible person	type	volume	constituent
contact	manufacturer	concentration	concentration
affiliation	dimension	molecular mass	pH

- the journal that cover topics related to food chemistry,

- the journal that do not have MIAPE-CC module as reporting guidelines requirement,
- articles submitted by researchers from Wageningen University,
- articles that are published in the time range from 2000 to 2014,
- articles that use column chromatography as a purification technique.

We deliberately excluded journals that explicitly require the compliance to the MIAPE-CC, since they might give a too positive impression of the use of terms from the guideline. We selected authors from Wageningen University as participants of the test group in our study. From 28 journals in total, specialized in food chemistry, 62 articles were retrieved. From these articles, we extracted the "Material and Method" sections – sometimes entitled "Experimental Method". Since the articles were retrieved in PDF format, we used Apache PDFBox [17] to parse and extract the method segments and stored them in plain text format. We created a CSV file including the title and the articles' DOIs. The collected method descriptions were marked as relevant by an expert from food chemistry domain. This set of method sections forms our corpus to evaluate the use of MIAPE-CC vocabulary. By counting how frequent each term occurs in the corpus, we can see how well the terminology required by MIAPE-CC is used by scientists. We used two packages from the RStudio toolbox for this experiment. The "tm" package was used to create the corpus and to pre-process the textual corpus. To have a more accurate mapping we transformed all tokens to the lower case. To prevent getting wrong mappings to commonly used words, we removed all stop words from our corpus [18]. The "qdap" package designed for quantitative discourse analysis was used to create a function – "termco" – to conduct the string mapping from our terms to the tokens [19]. The data and code are accessible through the Github (https://github.com/denatahvildari/MIAPE.git). The folder contains files related to the MIAPE-CC guideline, the developed vocabulary, the selected publications, and the R code used for the term occurrence experiment.

V. Result

The word occurrence measurement showed that from 83 terms in the vocabulary, 40 terms never occurred in any of the method description sections (48%). The 43 remaining terms occurred at least in one method section (51%). Table III provides the detailed results of the term occurrence experiment. The concept equipment contains 24 terms and it represents information about the product details for column, physical characteristics of column, and the chromatography system used for separation. From this class, 91% of the terms are not identifiable. Another interesting result is related to the concept 'column output'. Outputs of a run process are 'fraction' and 'detection'. Consider 'fraction' as an example. Descriptions about the start time and end time of fractionating process, and the size of fraction are essential information in this category. We observed that 55% of terms representing this concepts are not detected by our method. In the next section we discuss our observations and possible explanations for this result.

VI. Discussion

To gain some insight about this result, we consulted a domain expert and qualitatively analysed the data by inspecting the selected method sections. Our goal was not to find additional terms, but to identify generic patterns that explain the above results. We provide some explanations for these results. Only one term related to the MIAPE-CC category "general descriptors" occurred. The reason is that information about the name, contact, the date that the experiment was conducted, and the institutional role of the experimenter are not usually included in the "material and method" sections of publications. Information about the date is mostly documented in the scientists' laboratory notebooks. In the present model for publishing an article, this information can be found in the header along with the title of the paper. Authors do not see the necessity to report it in the method sections. This is common practice. The present underlying assumption is that this type of information is not assumed to be part of the experimental conditions needed for reproducibility.

TABLE III. TERM OCCURRENCE PER CATEGORY

Class	Never occurred terms	Occurred terms
General descriptor	4	1
Sample	9	8
Equipment	22	2
Mobile Phase	0	2
Column Run	0	5
Pre and Post Run processes	2	6
Column output	13	9

General information about samples and equipment such as name, manufacturer, model, and type is not detected by our method. The reason is that authors do not use the top level class terminology such as "manufacturer" to report the provenance of their experimental materials or equipment. They simply mention the name of the manufacturer. For example, consider the following sentence:

> "Branched sugar arabinon was obtained from British Sugar – Mcleary."

With our method, we searched for the term "manufacturer" and did not notice the fact that the British Sugar is an instance of the class 'manufacturer'.

Information about the mobile and stationary phases is crucial for describing a column chromatography experiment. However, the occurrence of these terms was not frequent in the selected 62 publications. As it is observed, the authors use synonyms when referring to the mobile phase, such as "solution", "eluent" and "solvent". For the same reason as described in the second observation, authors only mention the name of the mobile phase; for example, "solution (A): Water and solution (B): (ACN) Acetonitrile, Methyl cyanide". The term stationary phase was not frequently used. The stationary phase is the substance fixed in place for the chromatography procedure. In HPLC chromatography, the stationary phase is the same as the column and packing materials.

Terms representing the physical characteristics of the column were mentioned using abbreviations. For example, the inner diameter of the column is presented as "ID".

Information about the run processes are mostly mentioned along with information related to the column. In the MIAPE-CC model these concepts are categorised in separated classes.

The MIAPE-CC model indicates that for describing the column output, authors should mention the description about the detector that is used and how the fractionating procedure was done, if the experiment has gone through iterations. We could identify the detection equipment and some of the related terms such as the “wavelength”. However, the term “trace” never occurred. The reason is that a ’trace’ is being used for a specific type of a detector which is called “PDA”. In our selected publications this type of detector was never used.

VII. Conclusion and future work

In this paper, we argue that the reproducibility of an experimental method description is indebted to the existence of minimum information about that experiment. The minimum information guidelines specify all the details of an experiment such as materials, instruments, units of measure, characteristics of the column run processes and the possible deviations from the protocol. However, they are not highly adopted by researchers. This is partly because of the natural language nature of these guidelines, which does not allow for any computational support. This means that for reporting an HPLC experiment, a researcher still needs to follow extensive instructions and check too many lines to know which information is essential for describing a column run process. We believe that the existence of formal representations of these guidelines could improve their usage. We envision that if the vocabulary is applied in a software tool that scientists use on a daily basis, the transparency of the laboratory reports and consequently the reproducibility of the method can improve.

We investigated the MIAPE-CC guideline for reporting a column chromatography experiment and identified the main concepts and the associated terms. We evaluated its use in our domain of interest, which is food chemistry. Through an experiment we measured the occurrence of 83 terms from 7 categories in 62 method sections of published papers. The results indicate that half of the terms occurred at least in one of the descriptions. We mention that these results are not self-descriptive – meaning that the occurrence of terms does not guarantee the correct use of them, and also the absence of terms does not necessarily manifest the quality of the report. We realized this through a qualitative analysis and by consulting the domain experts. We learned that the our present method does not recognize the synonyms, abbreviations, instances and the existing relations. This is caused by the limitations in the model. Our analysis give a clear indication how to extend the vocabulary. With respect to its ultimate use, the present vocabulary is also limited in the sense that it does not present any semantic relations. These relations are needed when providing suggestions on missing information to authors when creating method sections or laboratory reports. We conclude that MIAPE is a good starting point for creating the required vocabulary, but it needs to be further elaborated. We should also mention that the sample size of the method description sections seems small (N=62), as some of the reviewers kindly pointed out, therefore results might not be conclusive. We take this remark into consideration for the next measurements. Nevertheless we see that even this small sample provided useful insight. The next step is to extend the vocabulary. For this, we use the Rapid Ontology Creation (ROC+) method. This tool is designed to be used by the domain experts, who do not have expertise in knowledge engineering. The method consists of two sessions, in which domain experts come together and jointly discuss, document and agree upon relevant terms and relations in their domain [20]. Moreover, we are looking into additional statistical methods to evaluate the mapping between the vocabulary and the method sections.

Acknowledgment

The authors would like to acknowledge the help of Prof. dr. Harry Gruppen and his team at Wageningen University.

References

[1] M. McNutt, “Reproducibility,” Science, vol. 343, no. 6168, 2014, pp. 229–229.

[2] E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber et al., “Promoting transparency in social science research,” Science (New York, NY), vol. 343, no. 6166, 2014, p. 30.

[3] J. B. Buckheit and D. L. Donoho, Wavelab and reproducible research. Springer, 1995.

[4] M. S. Anderson, B. C. Martinson, and R. De Vries, “Normative dissonance in science: Results from a national survey of us scientists,” Journal of Empirical Research on Human Research Ethics, vol. 2, no. 4, 2007, pp. 3–14.

[5] J. P. Ioannidis, M. R. Munafo, P. Fusar-Poli, B. A. Nosek, and S. P. David, “Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention,” Trends in cognitive sciences, vol. 18, no. 5, 2014, pp. 235–241.

[6] L. K. John, G. Loewenstein, and D. Prelec, “Measuring the prevalence of questionable research practices with incentives for truth telling,” Psychological science, 2012, p. 0956797611430953.

[7] A. Cabrera, W. C. Collins, and J. F. Salgado, “Determinants of individual engagement in knowledge sharing,” The International Journal of Human Resource Management, vol. 17, no. 2, 2006, pp. 245–264.

[8] C. Drummond, “Replicability is not reproducibility: nor is it good science,” 2009.

[9] A. De Waard, “The future of the journal? integrating research data with scientific discourse,” 2010.

[10] “The Minimum Information About a Proteomics Experiment (MIAPE),” 2010, URL: http://www.psidev.info/node/91 [accessed: 2016-04-13].

[11] “ARRIVE guidelines,” 2010, URL: https://www.nc3rs.org.uk/arrive-guidelines [accessed: 2016-04-13].

[12] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, “Improving bioscience research reporting: the arrive guidelines for reporting animal research,” Animals, vol. 4, no. 1, 2014, pp. 35–44.

[13] N. A. Vasilevsky, M. H. Brush, H. Paddock, L. Ponting, S. J. Tripathy, G. M. LaRocca et al., “On the reproducibility of science: unique identification of research resources in the biomedical literature,” PeerJ, vol. 1, 2013, p. e148.

[14] J. A. Medina-Aunon, S. Martínez-Bartolomé, M. A. López-García, E. Salazar, R. Navajas, A. R. Jones et al., “The proteored miape web toolkit: a user-friendly framework to connect and share proteomics standards,” Molecular & Cellular Proteomics, vol. 10, no. 10, 2011, pp. M111–008 334.

[15] P. E. Bourne, T. W. Clark, R. Dale, A. de Waard, I. Herman, E. H. Hovy, and D. Shotton, “Improving The Future of Research Communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331),” Dagstuhl Manifestos, vol. 1, no. 1, 2012, pp. 41–60. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2012/3445 - Retrieved on 16.03.2016

[16] C. F. Taylor, N. W. Paton, K. S. Lilley, P.-A. Binz, R. K. Julian, A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch et al., “The minimum information about a proteomics experiment (miape),” Nature biotechnology, vol. 25, no. 8, 2007, pp. 887–893.

[17] “Apache PDF Box,” 2010, URL: https://pdfbox.apache.org/ [accessed: 2016-04-13].

[18] I. Feinerer, “Introduction to the tm package text mining in r,” 2015.

[19] “Search For and Count Terms,” 2010, URL: http://finzi.psych.upenn.edu/library/qdap/html/termco.html [accessed: 2016-04-13].

[20] D. J. Willems, N. J. Koenderink, and J. L. Top, “From science to practice: Bringing innovations to agronomy and forestry,” Journal of Agricultural Informatics, vol. 6, no. 4, 2015, pp. 85–95.