The tissue microarray data exchange specification: Extending TMA DES to provide flexible scoring and incorporate virtual slides

Alexander Wright, Oliver Lyttleton, Paul Lewis, Philip Quirke, Darren Treanor

Department of Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, LS9 7TF, U.K., Institute of Life Science, School of Medicine, University of Wales, Swansea, SA2 8PP, U.K.

E-mail: Alexander Wright - a.wright@leeds.ac.uk

Received: 23 December 2010
Accepted: 15 February 2011
Published: 15 March 2011

DOI: 10.4103/2153-3539.78038

Abstract

Background: Tissue MicroArrays (TMAs) are a high throughput technology for rapid analysis of protein expression across hundreds of patient samples. Often, data relating to TMAs is specific to the clinical trial or experiment it is being used for, and not interoperable. The Tissue Microarray Data Exchange Specification (TMA DES) is a set of eXtensible Markup Language (XML)-based protocols for storing and sharing digitized Tissue Microarray data. XML data are enclosed by named tags which serve as identifiers. These tag names can be Common Data Elements (CDEs), which have a predefined meaning or semantics. By using this specification in a laboratory setting with increasing demands for digital pathology integration, we found that the data structure lacked the ability to cope with digital slide imaging in respect to web-enabled digital pathology systems and advanced scoring techniques. Materials and Methods: By employing user centric design, and observing behavior in relation to TMA scoring and associated data, the TMA DES format was extended to accommodate the current limitations. This was done with specific focus on developing a generic tool for handling any given scoring system, and utilizing data for multiple observations and observers. Results: DTDs were created to validate the extensions of the TMA DES protocol, and a test set of data containing scores for 6,708 TMA core images was generated. The XML was then read into an image processing algorithm to utilize the digital pathology data extensions, and scoring results were easily stored alongside the existing multiple pathologist scores. Conclusions: By extending the TMA DES format to include digital pathology data and customizable scoring systems for TMAs, the new system facilitates the collaboration between pathologists and organizations, and can be used in automatic or manual data analysis. This allows complying systems to effectively communicate complex and varied scoring data.

Key words: CDEs, DTD, tissue microarray, TMA DES, virtual pathology, XML

INTRODUCTION

Tissue microarrays (TMAs) are a low-cost, high-throughput technology, first described in 1998, for rapid analysis of protein expression and tissue morphology across hundreds of patient samples on a single slide. TMAs are typically made up of hundreds of cylindrical patient samples, known as cores, which are extracted from...
paraffin-embedded samples using a hollow needle. The cores are embedded in paraffin wax in a grid structure to be uniformly analyzed. Aside from the main benefit of dramatically increasing speed of analysis, another advantage of having a TMA in this arrangement is that they can be cut many times in order to be stained with different biomarkers for various immunohistochemical analyses.[2] Naturally, increasing the speed of analysis and creating multiple variations of the same set of tissue samples facilitates rapid growth of histological data. Having such an intensive data-generating technology requires organized and consistent storage methods in order to allow scalable and navigable data that can be used indefinitely.

Various technologies have been used to store TMA data in recent years, such as databases,[3,4] document object models,[5] and standardized spreadsheets,[6] all of which have their own advantages and disadvantages. A specifically web-based technology called the eXtensible Markup Language (XML) is also used in this field of research due to its concise, self-descriptive structure, and subsequent sharing capabilities and ease of use.[7]

The TMA data exchange specification (TMA DES) is an open source tool for describing, storing, and sharing TMA data, first described in 2003.[8] TMA DES has been designed using XML and ISO 11179 common data elements (CDEs) to allow TMA data to be stored using self-descriptive XML tags. The open source data format, utilizing the easily readable XML structure provides a lightweight, yet powerful, tool for facilitating TMA data transfer between collaborators and systems that use the published protocols. These protocols allow any persons or system using the same data structure to share their respective TMA resources independently from any proprietary- or platform-dependent systems.

Any given XML file can be validated against a set of rules relating to a specified data structure. This allows users and systems to identify whether the data they have conforms to their rules before trying to parse the data and use it. There are several ways of validating XML files, using XML schema languages, arguably the simplest to use being Document Type Definition (DTD) files.[9] The TMA DES has a freely available and ready-to-use DTD file for validating XML files that conform to its data structure.[10]

The TMA DES tool has been implemented in a variety of systems which either use the data structure as their core[11,12] or simply have the facility to import and export TMA DES valid XML files to provide a level of external compatibility.[3,4] We have chosen to utilize the TMA DES tool in our own virtual pathology TMA system, TMAi,[13] due to the well-documented protocols and potential for sharing data.

The increasing uptake of digital pathology over the recent years has led to digital TMA analysis becoming more common,[14] either manually through a computer screen instead of a traditional light microscope, or automatically, using specific advanced image processing algorithms. This imaging data has the capacity to be linked to pathological data, to facilitate storing data, results, and images all in one place.

However, whilst using the TMA DES data structure for our own TMA data over a period of several years, we found that we had no option for storing image locations or associated image data for virtual TMA slides. The lack of virtual slide data in the TMA DES structure is limiting for systems requiring their TMA data to be stored, accompanied by digital images of their TMA slides and individual cores with associated data. This limitation curtails the ability to perform large-scale image analysis automatically, as the defined core images are not addressable. Second, when scoring cores on a TMA, we found the core results CDEs limiting in what information we could store about the cores. It is these limitations which we propose to mitigate by extending the TMA DES CDEs for use with digital pathology systems.

**MATERIALS AND METHODS**

The existing TMA DES data structure requires four main data types for an XML file to be valid. The CDEs that describe an overview of the TMA data (<header>), the TMA blocks (<block>), the slides cut from that block (<slide>), and the cores that make the block (<core>) are all mandatory elements that must be present in order for the DTD to validate a given XML file. The nesting of the tags reflects which elements belong to which, for example, a slide that has been cut from a TMA block will be modeled by nesting under the appropriate block element. Figure 1 shows an example of an empty, yet still valid XML data structure with appropriate nesting.

In order to extend the TMA DES to include digital pathology data, we will be focusing on the slide and core CDEs, which will allow us to incorporate virtual slide images, for both whole slides and individual cores. We will then look at extending the data exchange specification to incorporate more flexible scoring methods.

**Part I: Digital Pathology Image Data**

We propose to extend the slide CDEs by adding a child element that contains the virtual slide’s location, using a uniform resource indicator (URI), which can be used to look up and view the slide for which the data are being stored. This can be a URL of a web-hosted slide, a Digital Object Identifier, or any other appropriate slide image file. The additional CDE is shown in Figure 2 in bold type. A full list of the proposed extended CDEs is listed in Appendix A, in ISO/IEC 11179 format.
Note that in this case, the slide URI should be an overall view or thumbnail image of the entire TMA. This will enable any user of the data to check the virtual slide or automatically list the entire collection of slide images contained in the XML file, in order to navigate the data visually.

We also propose to extend the core and its child elements in a number of ways, also because of limitations in describing virtual slide data. First, the core CDEs are extended by adding location data CDEs, highlighted in Figure 3.

The location data for each core is essential for handling core data, where there are cores on the same slide using the same array identifier, as well as being fundamental to imaging tools that match data to core locations. The core grid element provides an option to identify a TMA, where more than one TMA is embedded in the same paraffin block. The row and column elements express the core position in an integer value for its respective placement, and the x and y coordinates relate to the position of the core in mm, based upon the TMA block design. As with the <slide_uri> CDE, there is a full description of the CDEs in Appendix A.

Part II: TMA Results

Nested under the core CDE are more elements, which describe data related to the core. These describe the core image data and the core results data [Figure 4]. In order to extend the data exchange specification to work with virtual pathology, we have looked at the core image and, in more detail, the core results CDEs.

The existing core image CDE contains information about the image of the core itself, such as magnification and format, but specifies no information about where the core is stored. As with the slide CDE, an element to store a URI of a core image or virtual slide (with coordinates to point to the specific core) has been added to allow the actual core images to be linked to the core data, along with the image data itself.

As it is well established that TMAs are used for high throughput of data,[1,2] it is sensible to focus on the results of TMA scoring as the most important reason for storing such data digitally. Used in practice, we found
that the TMA DES CDEs did not encapsulate the varied scoring styles and techniques employed by pathologists even within our own department, and found that having a static set of scoring category CDEs was too limiting for scoring cores using different biomarkers, stains, and methods. The existing CDEs provide data storage for predefined fields, which are limited to numeric values of tissue intensity and percentage of tissue staining, and give no indication of what part of the core is being scored, be it nuclear, cytoplasmic, membranous, etc.

To extend the data exchange specification into something more usable and more generic, we first looked at how a core could be scored, and what data were kept about the scores. Then, we looked at how this could fit in with the existing CDEs to provide a minimal amount of extra elements to add to the specification.

Figure 6 shows an illustrated example of how a single core in a TMA can be given multiple scores, which can be chosen from categories a pathologist has devised specifically for a project, percentages, numeric or text classifications, or more generally used categorizations such as nuclear intensity. Each scoring system can have as many subscores as considered suitable for the purpose, and may vary greatly depending on factors such as tissue, disease, stain, and purpose of scoring, to name a few. The variety in which a single core is scored in practice requires a more robust and adaptable scoring data structure for storing these valuable results. The diagram illustrates that the scores are meaningless without proper semantic definition.

When observing pathologists using different scoring systems, it became apparent that there is a uniform way of being able to describe these systems without being restrictive, and also without being overly complex and creating an unnecessary amount of data fields. Table 1 shows three different scoring systems; X, Y, and Z, which have been broken down into a number of subcategories, and each subcategory is then broken down into a collection of scoring categories or percentages.

From Table 1, we can clearly see that the scores available to the pathologist in each scoring system are described in different ways, which implies that the scoring systems used in TMA analysis must be described in the TMA data exchange specification, as well as the score for that scoring system. We propose to extend the TMA DES structure for scoring by adding in a small set of CDEs which allow the definition of a scoring system, the subcategory of the scoring system being used, and the score itself. Also, the addition of a scorer has been prepended to this list of new CDEs, in order to attribute work, and assist interobserver variations in scoring. This is particularly important when more than one observer has to score a TMA, for example, in biomarker validations.

Figure 7 illustrates the new CDEs, and how they fit into the core results CDE. Note that by simply adding in these five new elements, we can accurately describe a core in relation to any given scoring system, what the score actually means, the scoring system itself, and who the core has been scored by. Using this method, we also allow a core to be scored multiple times with either the same or different scoring systems. Figure 8 gives a full example of a core result CDE using the extended elements, and data from Table 1.

Figure 8 illustrates that by using the new format, the XML structure is now capable of storing multiple scores for multiple scoring systems, which can be user defined and completely unique.

In order to ensure that the new format was valid, we first ran the XML through a validator using the original TMA DES DTD.[10] Expected errors occurred, showing that
the new CDEs were not accepted by the current DTD. After testing the new XML format on the original DTD, we proceeded to make a new DTD file which could successfully identify the new CDEs and validate the new format correctly. The DTD file was created using the CDE descriptions in Appendix A.

The TMA DES DTD format allows for extension of the original CDEs by providing an external DTD that can also be referenced. This file contains Local Data Elements (LDEs) which can be defined by anyone wishing to validate an extended version of the specification to fit their own laboratory standards. The extensions we are proposing are suggestions for all users of TMA DES to incorporate, so we are primarily extending the main DTD file of CDEs, rather than a local LDE doctype file. We have, however, produced both extended CDE and LDE DTD supplementary files to use as required.

**RESULTS**

Using our own TMA database system, built on the TMA data exchange specification,[13] a test TMA Project dataset was used to create a TMA DES XML file containing data for 3 TMAs, 40 slides, 492 cores, 6708 images, and 4,652 scores.

First, the file was tested using the new DTD files of both extended CDEs and LDEs to validate it, and invalid XML files containing misspelled variations of the new elements were also passed to the DTDs. The validation results showed that the new DTDs handled the new data correctly, and that non–well-formed and invalid data were not accepted in all cases.

**Table 1: Examples of scoring systems broken down into sub categories**

| Scoring system | Sub category          | Score | Description                                           |
|----------------|-----------------------|-------|-------------------------------------------------------|
| Scoring system X | Positivity            | 0     | No staining present                                   |
|                |                       | 1     | Staining present                                     |
| Scoring system Y | Nuclear percentage    | (Percentage) | Percentage of nuclear staining                      |
|                | Nuclear intensity     | 0     | No staining                                           |
|                |                       | 1     | Weak staining                                         |
|                |                       | 2     | Moderate staining                                     |
|                |                       | 3     | Marked staining                                       |
|                | Cytoplasmic percentage| (Percentage) | Percentage of cytoplasmic staining              |
|                | Cytoplasmic intensity | 0     | No staining                                           |
|                |                       | 1     | Weak staining                                         |
|                |                       | 2     | Moderate staining                                     |
|                |                       | 3     | Marked staining                                       |
| Scoring system Z | Nuclear % staining    | 0     | No positive nuclei                                    |
|                |                       | 1     | Between 0% and 33% positive nuclei                  |
|                |                       | 2     | Between 33% and 66% positive nuclei                  |
|                |                       | 3     | Over 66% positive nuclei                             |
|                | Cytoplasmic staining  | 0     | No cytoplasmic staining                               |
|                |                       | 1     | Cytoplasmic staining present                         |

Figure 8: An example result using scoring system X from Table 1
To validate the effectiveness of the new CDE format, we used our exported dataset to (a) use with image analysis algorithms and (b) to review multiple observer scoring results for comparison.

Figure 9 shows the benefits of having the core image URI stored with the core data, as the XML data show the core image data (A), which can then be looked up and reviewed manually (B), or can be passed automatically using image processing algorithms for further analysis (C). A further benefit of having the human scores stored with the URIs is that the human scores can also be used to validate imaging algorithms either being developed or trained on TMA datasets.

Also shown in Figure 9 (A) the benefit of storing the scoring data in this way is that interobserver variability can be easily calculated on desired scoring systems for the same slide or core. The XML in Figure 9 shows scores for the pictured core (C) using an immuno-scoring system scored by multiple observers. The table in Figure 9 (B) shows the results in a clearer format. The XML also highlights again the use of multiple scoring systems and benefits of subcategorizing scoring systems for scoring different features of the cores, whilst keeping the data identifiable to a given classification.

**DISCUSSION**

From using the TMA DES format extensively in our work, it has become apparent that there is a need to be able to express scores in a nonrestrictive way that is appropriate to a specific project, tissue, stain, or simply a particular observer. It is also important to make this data format easy to use and understand, and allow reuse of these standards to enable the best possible sharing of data with multiple organizations.
The increasing use of digital pathology resources and image analysis software provides a perfect platform for combining with digitized TMA data. Our new proposed extension of the TMA data exchange specification facilitates the use of TMA images, for automated or manual analysis at both slide and core level, and also supports any type of applicable scoring data, providing a collaborative scoring tool between pathologists with custom scoring systems. We believe that the intricate scoring systems and variability of scoring between pathologists, organizations, and clinical trials was not provided by the existing TMA DES standard, and are not aware of any other tools providing such functionality.

Scoring systems exist which are standardized, but the majority of scoring systems used in clinical trials are ad hoc and based significantly on the exact clinical question being asked. The proposed extension to TMA DES has an advantage in that it allows the sharing of existing scoring systems, thus improving agreement between studies while also providing the facility to extend or invent new scoring systems. By having the TMA and core images available to anyone using the extended scoring CDEs, core data can easily be added or reviewed based upon the visual information from the readily available images, specified by their URIs. An alternative to extending the TMA DES CDEs is to store the TMA image and scoring data separately, but this would reduce interoperability by keeping data apart, and requiring more than one simple XML data file.

Our new TMA DES extension has been implemented successfully in our own web-based TMA system, which stores data across multiple TMA projects, containing over 37 000 cores, each with multiple observers and scoring systems.[13] With this data, we have used the new core image URI CDE extensively to use with automated image analysis, and also for developing our own image analysis algorithms, and evaluating them on large datasets.[15] A significant advantage to this approach is that large datasets such as these can be shared with external investigators for clinical trials or biomedical research, who can then use the TMA data in for their own analyses or validation. The proposed extensions to the TMA data exchange specification are readily available for any groups wishing to use this functionality at http://www.virtualpathology.leeds.ac.uk/tmade.

By extending the data exchange specification to include a small set of new CDEs that incorporate elements for use in collaborative digital pathology, we have enhanced the TMA DES protocol by enhancing interoperability between digital TMA scoring repositories, and also by encouraging real-world use of TMA DES systems to score, analyze, and store data.

| Slide URI | Identifier: slide_uri | Version: 1.0 | Language: EN-UK | Obligation: Optional | Datatype: Character String | Maximum Occurrence: Unlimited | Definition: A URI that links to the virtual TMA slide | Comment: The URI can point to an image of the slide, which may be the full image, a static thumbnail image, or a multi scale image viewing component, hosting the slide full image (e.g., zoomify). |
|-----------|-----------------------|-------------|-----------------|-----------------------|-----------------------------|-----------------------------|-------------------------------------------------|-----------------------------------------------|
| Core Row  | Identifier: core_row  | Version: 1.0 | Language: EN-UK | Obligation: Optional | Datatype: Integer          | Maximum Occurrence: Unlimited | Definition: A number denoting the core row position in the TMA block | Comment: |
| Core Column| Identifier: core_column| Version: 1.0 | Language: EN-UK | Obligation: Optional | Datatype: Integer          | Maximum Occurrence: Unlimited | Definition: A number denoting the core column position in the TMA block | Comment: |
| Core Grid | Identifier: core_grid | Version: 1.0 | Language: EN-UK | Obligation: Optional | Datatype: Character String | Maximum Occurrence: Unlimited | Definition: An identifier that groups cores from the same TMA block, where there are multiple TMA blocks on the same slide | Comment: |
| Core X Co-Ord| Identifier: core_x_co-ord | Version: 1.0 | Language: EN-UK | Obligation: Optional | Datatype: Real Number | | |
| Identifier         | Description                                                                 |
|--------------------|-----------------------------------------------------------------------------|
| core_y_co-ord      | A number expressing the position of the core in the TMA block along the y axis in mm |
| core_image_uri     | A URI that links to the image of the core from the virtual TMA slide        |
| core_results_scoring-system | The name of the scoring system used to score the core                   |
| core_results_scoring-system_score | A score relating to the parent core_image and being a specific value from the chosen core_results_scoring-system_sub-category |
| core_results_scoring-system_sub-category | This is an overall name for the scoring system (e.g., “Project A Immuno Scoring System”), as it can contain multiple specific scores (see core_results_scoring-system_sub-category) |

**REFERENCES**

1. Kononen J, Bubendorf L, Kallioniemi A, Bärlund M, Schraml P, Leighton S, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 1998;4:844-7.
2. Kallioneimi OP, Wagner U, Kononen J, Sauter G. Tissue microarray technology for high-throughput molecular profiling of cancer. Human Molecular Genetics 2001;10:657-62.
3. Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, et al. The
Standford Tissue Microarray Database. Nucleic Acids Res 2007;36:D871-7.

4. Thallinger GG, Baumgartner K, Pirklbauer M, Uray M, Pauritsch E, Mehes G, et al. TAMEE: Data management and analysis for tissue microarrays. BMC Bioinformatics 2007;8:81.

5. Lee HW, Park YR, Sim J, Park RW, Kim WH, Kim JH. The Tissue Microarray Object Model: A Data Model for Storage, Analysis, and Exchange of Tissue Microarray Experimental Data. Arch Pathol Lab Med 2006;130:1004-13.

6. TMA-TAB: A spreadsheet-based document for exchange of tissue microarray data based on the tissue microarray-object model. J Biomed Inform 2010;43:435-41.

7. Bray T, Paoli J, Sperberg-McQueen CM. Extensible Markup Language (XML) 1.0. Available from: http://www.w3.org/TR/1998/REC-xml-19980210.pdf [Last accessed on 2010 Aug 05].

8. Berman JJ, Edgerton ME, Friedman BA. The tissue microarray data exchange specification: A community-based, open source tool for sharing tissue microarray data. BMC Med Inform Decis Mak 2003;3:5.

9. Lee D, Chu WW. Comparative Analysis of Six XML Schema Languages. SIGMOD Record 2000;29:76-87.

10. Nohle DG, Ayers LW. The tissue microarray data exchange specification: A document type definition to validate and enhance XML data. BMC Med Inform Decis Mak 2003;5:12.

11. Sharma-Oates A, Quirke P, Westhead D R. TmaDB: A repository for tissue microarray data. BMC Med Inform Decis Mak 2005;6:218.

12. Berman JJ, Datta M, Kajdacsy-Balla A, Melamed J, Orenstein J, Dobbin K, et al. The tissue microarray data exchange specification: Implementation by the Cooperative Prostate Cancer Tissue Resource. BMC Med Inform Decis Mak 2004;5:19.

13. Wright A, Lyttleton O, Lewis P, Quirke P, Treanor D. TMAi | An open source Tissue Microarray database using published XML standards. Biomedical Informatics without borders: From collaboration to implementation. A joint conference of the U.S. National Cancer Institute and the U.K. National Cancer Research Institute Informatics Initiative. London, UK: Wellcome Trust; 2009. p. 41-2.

14. Rojo MG, Bueno G, Słodkowska J. Review of imaging solutions for integrated quantitative immunohistochemistry in the Pathology daily practise. Folia Histochemet Cytobiol 2009;47:349-54.

15. Wright A, Magee D, Quirke R, Treanor D. Automated scoring of Tissue Microarrays using virtual slides. Joint Meeting of the Pathological Society of Great Britain and Ireland and the Dutch Pathological Society. UK: University of Leeds; 2008. p. 64.