



TÜRK STANDARLARI ENSTİTÜSÜ
TURKISH STANDARDS INSTITUTION

COMMENTS
ON THE STANDARD
ISO/IEC DIS 29500 (ECMA 376:2006)
OFFICE OPEN XML FILE FORMATS

Summary:

This report has been prepared for justification of the vote of Turkey for the draft standard of, ISO/IEC DIS 29500 (ECMA 376:2006) Office Open XML File Formats standard, which is prepared by ECMA.

Prepared By: ICT Expert group of Turkish Standards Organization

Dr. Ali ARİFOĞLU (Chair)

Asst. Prof. Dr. Mehmet Demirer (Deputy Chair)

Gökhan Şengül (Member)

Osman Öz (Member)

July 2007
Ankara



Table of Contents

1	INTRODUCTION	3
1.1	SCOPE	3
1.2	ABBREVIATIONS	3
1.3	REFERENCES	3
1.4	OPEN FORMATS	4
2	XML AND OTHER RELATED STANDARDS	5
2.1	XML	5
2.2	OFFICE OPEN XML – ISO/IEC DIS 29500	6
2.3	ODF -ISO/IEC 26300	10
3	DISCUSSIONS ON ODF AND OPEN XML	11
3.1	ODF	11
3.2	OPEN XML	13
4	CONCLUSION	15



1. Introduction

1.1. Scope

This report has been prepared for commenting on the draft standard ISO/IEC DIS 29500 [5] of which voting process is going to end at the beginning of September 2007. The standard basically includes the description of office documents by XML which is a standard specification language. This standard and the similar ones are specifically important for the interoperability of organizations. The report is prepared by the ICT Expert Group of ISO.

In the report, the concepts of open formats, XML, ODF and Office Open XML are explained first, and then, the discussions on the above mentioned standard and rationale for the voting were explained.

1.2 Abbreviations

ANSI	: American National Standards Institute
AVI	: One of audio file compression formats
ICT	: Information and Communication Technology
BMP	: One of picture file compression formats
ECMA	: European Computer Manufacturers Association
GML	: Generalized Markup Language
HTML	: HyperText Markup Language
IEC	: International Electrotechnical Committee
IPR	: Intellectual Property Rights
ISO	: International Standards Organization
ODF	: Open Document Format
OOXML	: Office Open XML
SGML	: Standardized Generalized Markup Language
TSO	: Turkish Standards Organization
W3C	: World Wide Web Consortium
XML	: eXtended Markup Language
Y2K	: year 2000 problem
ZIP	: One of file compression formats

1.3 References

- 1 A Brief History of the Development of SGML,
<http://www.sgmlsource.com/history/sgmlhist.htm>
- 2 **Ecma international**. TC45 - EXPLANATORY REPORT ON OFFICE OPEN XML STANDARD (ECMA-376) SUBMITTED TO JTC 1 FOR FAST-TRACK
- 3 http://en.wikipedia.org/wiki/Office_Open_XML
- 4 <http://www.forumex.net/asp-perl-php-html/35649-xml-ve-xml-uygulamalari.html>
- 5 ISO/IEC 26300:2006 Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0
- 6 ISO/IEC DIS 29500 (ECMA 376:2006 Office Open XML File Formats)
- 7 Mark Johnson, XML for the absolute beginner,
<http://www.javaworld.com/javaworld/jw-04-1999/jw-04-xml.html>
- 8 Micheal Morrison et al., *XML Unleashed*, Sams Publishing, 1999.
- 9 Ralf I. Pfeiffer, Tutorial 1: Overview of XML,
<http://www4.ibm.com/software/developer/education/tutorial-prog/overview.html>
- 10 Richard Anderson et al., *Professional XML*, Wrox Press Ltd., 2000.



1.4 Open Formats

At the beginning of the millenium, the countries of the world who had solved their Y2K problems, accelerated their e-Transformation process. This resulted with the exhaustive usage of internet all over the world. Over next decade, the commerce is expected to move on internet and the world is going to a life dependent on ICT with high speed. The countries who are well developed on ICT usage are also changing their economy into “knowledge economy”too. The societies are changing to knowledge society while classical economies are transforming to knowledge economy.

Undoubtedly, the basic infrastructure for e-transformation is ICT. The governments and organizations should expand their ICT usage as much as they can and they should reorganize their busines processes accordingly. Business environments are changing to “paperless offices” while arcives are being replaced by electroinic counterparts. The office software such as word processes, spreadsheets, presetanation and database applications are becoming a natural part of daily business life.

Today, for all ICT applications, it is almost compulsory

- To produce documents automatically from the data stored in data bases and similar media,
- To extract the information within the current documents and to store them into structured media such as data bases,
- To run on a variety of hardware devices such as mobile equipments etc.,
- To work with applications in different environments such as in other organizations in harmony, which means talking the same language with them (**interoperability**).

This obligation has resulted with the development of XML¹ (eXtended Markup Language) which is accepted by all parties withot any hesitation. XML which is developed by W3C (World Wide Web Consortium) became a unique language for data interchange and substituted with classical EDI. Today, the information and services given on web sites are all provided via XML.

Open XML formats needs to be standardized because of following major reasons: :

- the ability of processing of binary information,
- technological develoments,
- the pressure of market applications,

¹ www.xml.org, www.w3.org/TR/REC-xml/



- the need of conversion and long term storage of billions and billions of existing documents.

2. XML and related standards

2.1 XML

XML is the Extensible Markup Language in expansion. XML is a basic and flexible text formatting technology, which can be used as a strategic tool in the field of electronic commerce, electronic data interchange, data management, and search engines motors. Structures, contents and concepts of data can be represented in stand-alone way, without having dependencies on a platform, a company and a language.

By the development of press publications, the notes and special symbols that publishers prepare to mark in the press machines are represented as “markup”. This is a process of marking to emphasise a certain parts of texts. The marks, rules and grammar info sets used are defined as a “markup language”

The word processor programs contain many marks embedded into the text to parse types, parts and styles. Programming languages use several symbols and marks to parse functions, data structures and data. Without using this kind of a parser, a mark or a set of tags, there is no possibility that an application can be developed [4].

The first markup language GML (Generalized Markup Language) was first used in the end of 1960 to transfer, share and process texts and documents as a result of research in IBM. The GML was then improved by a group set up in ANSI (American National Standard Institute) in 1978 and was accepted by ISO (International Organization for Standardization) as a Standard named SGML (Standardized Generalized Markup Language) in 1986. SGML is a language that determines a language’s semantics and grammar used in text and documents sets. SGML is already used as a documentation Standard in the organisations of US government, avionics, automotive and press industry. SGML cannot be used commonly due to its high developing and application cost as well as its highly complicated structure despite it is a very powerful language [1].

In 1989, Tim Berners-Lee and Anders Berlung developed HTML (Hypertext Markup Language) to easily share documents through the internet, which is one of the fundamental elements of web applications. HTML was developed as an application of a SGML. In other words, the structure of the HTML language has been defined in the SGML. HTML is a language to display and format information in a Standard form, such as a header, a script type, a picture and a table in computer environment. The presentation of a document is realised through the use of several marks called tags. The main purpose to develop this language is to display and present a document in a Standard form. The development of this language for only web browsers, as well as many other restrictions have caused the XML language to be developed afterwards [7].



In 1996, the World Wide Consortium (W3C, <http://www.w3.org>) started to design the XML Language as a simple markup language, in order to include the strength and flexibility of SGML. In February 1998, XML 1.0 was published by the W3C as a Standard. XML is a simplified language, including many features of the SGML Language and is a subset of the SGML. XML is a meta language just as the SGML. In other words, it is used to define structures of other languages [9].

XML is also a language that uses the tags as the HTML does. The main difference between the HTML and the XML is that the tags are used to define the contents of information. The XML is a meta language. It is used to define the other markup languages.

Using XML, an application special markup language can be defined for any application to express the content of data and data types. Metadata or Metainformation is information about data. The XML tags define meta data about data [8].

The XML introduce an appropriate medium to define and represent various data, concept and contents. For this reason, the XML becomes widespread rapidly as a strategical tool for definition and transfer of application data in various areas, without depending on producer, language and platform. Some of the main application areas in which the XML has been used and to be used are given below [10]:

- Internet search engines,
- Device and application independent data communication,
- Electronic commerce,
- Software development,
- Press and Publishing

2.2 Office OPEN XML - ISO/IEC DIS 29500

The Office Open XML (commonly abbreviated as OOXML) is a file format specification for electronic documents such as word processing documents, presentations, charts, spreadsheets, books and reports. The OpenXML is an open standard draft and it can be freely implemented by multiple applications on multiple platforms. The major benefit of publication of this standard is stated as provision of a common platform for all organizations developing application software, an the entities using those software as well as for the educators or authors who teach the format [3].

The work to standardize Open XML has been carried out by Ecma International via its Technical Committee 45 (TC45), which includes representatives from Apple, Barclays Capital, BP, The British Library, Essilor, Intel, Microsoft, NextPage, Novell, Statoil, Toshiba, and the United States Library of Congress



[3].

Office Open XML format uses a ZIP container for packaging XML and other data files. The main advantage of Open XML is backward compatibility and it supports the files created before the Open XML format. It has been declared that the Open XML standard draft meets the European Union definition of an Open Standard.

The Open XML is a file system that contains the individual files that form the basis of a document. In addition to XML files the ZIP package can also include binary files in formats such as PNG, BMP, AVI or PDF.

Open XML was designed from the start to be capable of faithfully representing the pre-existing corpus of word processing documents, presentations, and spreadsheets that are encoded in binary formats defined by Microsoft Corporation. The standardization process consisted of mirroring in XML the capabilities required to represent the existing corpus, extending them, providing detailed documentation, and enabling interoperability. At the time of writing, more than 400 million users generate documents in the binary formats, with estimates exceeding 40 billion documents and billions more being created each year [6].

Concurrently with diversified marketing, a new range of applications not originally contemplated in the document editing programs are introduced. These new applications include ones that:

- generate documents automatically from business data;
- extract business data from documents and feed those data into business applications;
- provide accessibility for user populations with specialized needs; or
- run on a variety of hardware, including mobile devices.

It is declared that this standard draft has the capability of long-term preservation. In parallel with the developments in science and related fields we have learned to create exponentially increasing amounts of information. Those information has been encoded using digital representations that are deeply coupled with the programs that created them after a decade or two, they routinely become extremely difficult to read without significant loss. Preserving the financial, scientific, intellectual and other related investment in those documents (both existing and new) has become a pressing priority [2].

It is declared that there are four main reasons to introduce the Open XML standard draft: extremely broad adoption of the binary formats, market forces



hat demand diverse applications, technological advances, and the increasing difficulty of long-term preservation. These reasons cause the not only the development of the draft, but also they cause the migration of billions of documents to it with as little loss as possible. On the other hand, standardizing that open XML format and maintaining it over time create an environment in which any organization can safely rely on the ongoing stability of the specification, confident that further evolution will enjoy the checks and balances afforded by a standards process [2].

Various document standards and specifications exist; these include HTML, XHTML, PDF and its subsets, ODF, DocBook, DITA, and RTF. Like the numerous standards that represent bitmapped images, including TIFF/IT, TIFF/EP, JPEG 2000, and PNG, each was created for a different set of purposes. Open XML addresses the need for a standard that covers the features represented in the existing document corpus. It is declared that it is the only XML document format that supports every feature in the binary formats [3].

OpenXML defines the formats for word processing, presentation and spreadsheet documents. Each document type is specified by one of the WordprocessingML, PresentationML or SpreadsheetML markup languages.

Some Features of OpenXML Draft Standard Draft

In this section some of the important features of OpenXML are given.

1. **Interoperability:** Developers can write applications that consume and produce OpenXML on multiple platforms. Foremost, the interoperability of OpenXML has been accomplished through extensive contributions, modification, and review of the Specification by members of the Ecma TC45 committee with diverse backgrounds and corporate interests. Representation included:
 - Vendors (Apple, Intel, Microsoft, NextPage, Novell, and Toshiba) with multiple operating systems (Linux, MacOS, and Windows) and multiple intended uses of OpenXML
 - Corporations (BP, Barclays Capital, Essilor, Statoil) with heavy investments in existing content, including mission-critical transaction systems
 - The British Library and the United States Library of Congress, both of whom have direct interest in preservation.
2. **Internationalization:** OpenXML supports internationalization features required by such diverse languages as Arabic, Chinese (three variants), Hebrew, Hindi, Japanese, Korean, Russian, and Turkish. OpenXML inherently supports Unicode because it is XML. In addition, OpenXML has a rich set of internationalization features that have been refined over the course of many years.



3. **Low barrier to developer adaption:** An experienced developer can begin to write simple OpenXML-conformant applications within a few hours of beginning to read the Specification. Although the specification describes a large feature set and the standard draft is very long, an OpenXML-conformant application need not support all of features in the Specification.

4. **Compactness:** The Open XML file format supports the creation of high-performance applications. The main reason for it is that an Open XML file is conventionally stored in a ZIP archive for purposes of packaging and compression, following the recommended implementation of the Open Packaging Conventions. As a result Open XML files are on average 25% smaller, and at times up to 75% smaller, than their binary counterparts. A second simple source of compactness, particularly where an uncompressed representation is required, is the length of identifiers in the XML. Frequently used tag names are short. Implementers are encouraged to use short namespace prefixes.

5. **Modularity:** An application can accomplish many tasks by parsing or modifying a small subset of the document. Three features of the Open XML format cooperate to provide this modularity.
 - A document is not monolithic; it is built out of multiple parts.
 - Relationships between parts are themselves stored in parts.
 - The ZIP archive format that is typically used to support Open XML documents supports random access to each part.

6. **Integration with business data:** Open XML enables organizations to integrate productivity applications with information systems that manage business processes by enabling the use of custom schemas within Open XML documents. An organization's goals in taking this approach would be to reuse and to automate the processing of business information that is otherwise buried opaquely inside documents, where business applications cannot read or write it.

7. **Room for innovation:** Open XML is designed to encourage developers to create new applications that were not contemplated when the binary formats were defined, or even when Open XML was defined [2].



2.3 ODF - ISO/IEC 26300

The Standard “ISO/IEC 26300: Information technology – Open document format for office applications (OpenDocument) V1.0” about open formats was published by ISO in 2006. OpenDocument is the abbreviation of OASIS Open Document Format for Office Applications and also known as ODF. ODF is a document file format that it is used to define notes, reports, books, electronic tables, schemas and word processor files. This Standard is developed by a Technical Committee under the consortium of “Organization for the Advancement of Structured Information Standards” and is based on the XML format first developed and implemented by the OpenOffice.org developer of Office Applications. The Standard ODF can be freely obtained and used. Therefore, this Standard satisfies all fundamental definitions that an open Standard has to employ. In other words, a software developer can learn about details of this format and can develop application softwares that can read these files and that can produce files in this form. This file format is also the format that the applications such as OpenOffice.org 2.0, KOffice 1.5, StarOffice 8, IBM Workplace etc use. The Standard ISO/IEC 26300 published by ISO (International Organization for Standardization) in May the 1st, 2006 has formed a file format that can be used worldwide to store files produced by the Office applications. This Standard is also the first Standard in the world in its area. The users of software are guaranteed that they can use their data now and in the future, using an appropriate software package. This means that any open Standard compatible application can use it.

The Standard ISO/IEC 26300:2006 defines XML schemas and semantics for the Office applications. The schema defined here can be implemented for the Office documents such as text documents, electronic tables, schemas, drawings and presentations. However, the area of applications is not limited by the examples mentioned above.

The Standard ISO/IEC 26300: 2006 provides advanced-level information for organisations of documents. It also defines appropriate structures of XML for the Office documents and is also convenient for conversions to be performed by using XML based or similar tools.

The Standard ISO/IEC 26300: 2006 primarily provides beginners level information for the OpenDocument format and explains the structure of documents satisfying OpenDocument specifications. It also presents meta data of these documents and Ayrıca bu dokümanlarda yer alabilen meta information and paragraph and text contents of this information.

ISO/IEC 26300:2006 defines the content table of a document of the OpenDocument format, its graphical content, schema content and content format. It also defines a common content info for all documents [5].



3 Discussions related to ODF and Open XML

3.1 ODF

The biggest proof that open document format standard evolution continues is because these studies are regularly criticized and the need to renew and update them is apparent. It is also obvious that the criticism related to these formats will continue for a long time. When we look at the ODF standard, we see the following basic criticisms about it:

- **ODF ISO specification does not contain a defined formula language.** This means that ISO conforming files do not have to be compatible. There are works on this issue for the new version of ODF.
- **There is no macro language specified in ODF and this results in different methods of providing macro/scripting capabilities in different applications using ODF.** Many things can be said about this and it is also debatable whether or not a macro language is necessary but office documents somehow contain macros and we think that the standard also should contain them.
- **ODF ISO specification does not allow for tables in presentations.** There are temporary solutions to this and the need for a permanent solution is obvious. There are works on this issue for the new version of ODF.
- **While ODF is aimed to be an open format standard, it supports many features of OpenOffice but almost none of Microsoft Office.** There are two points in this case: 1. Microsoft is punished because of not joining the preparations of the standard. 2. These features are not in the scope and target of the standard and there is a need for another standard with different requirements and definitions.
- **ODF does not support digital signatures.** Open Office allows digital signatures and stores them in ODF files, but ODF specification does not contain any detail related to digital signatures. Digital signature capabilities are expected to be present in the new version of ODF.
- **ODF is not using existing ISO standards and commonly used formats for mathematical formulas.** ODF uses W3C standard MathML for mathematical formulae but there are some formula presentation problems in MathML. Therefore, ODF is criticized for not using ISO 12083 and TeX which is commonly used.
- **Java applets are described as native objects in the ODF specification.** The problem here is that any full implementation will require a Java VM present from within the application while conformance to the standard does not require a full implementation.



- **The ODF standard is insufficiently detailed.** This requires the creation of excessive application specific namespace extensions to record document features. This issue is one of the tradeoff points for any XML based specification. Either it will be more detailed and will result in a large specification or it will be limited and the result will be a smaller specification. The former will compromise the flexibility but will prevent namespace complexity while the latter will result in namespace complexity.
- **There is a precondition about Intellectual Property Rights (IPR) and patent issues.** There are concerns about the Sun Microsystems' patent (openness) statement related to future versions. Sun Microsystems declared that this statement is valid only if they participate in the development of any future version and this creates an uncertainty about openness.



3.2 Open XML

For a group of documents with different requirements and where ODF is not able to find solutions, another open document format standard (Open XML) is being prepared. ISO/IEC DIS 29500 is still a draft standard and there are many ongoing discussions and criticisms' related to this work.

Open XML criticisms reached to a point where technical issues are less important than the commercial and political ones. The following are some of the important criticisms related to this draft standard:

- **There is already an open document format therefore there is no need for a second ISO standard for the same purpose.** While it may look like correct theoretically, it is not correct in practice. There are different standards for different image compression and document types. While their purposes are same, their targets are different. The targets and target document types of the ODF standard are different than the ones for Open XML. Because of this structure, it is not a problem to have two, even more, standards based on XML with the same purpose but addressing different requirements.
- **The size of the standard is big (6000 pages) and therefore standardization process must be slower, not fast track.** Like ODF, Open XML mainly consists of identifiers and elements comprising the specification. Compared to ODF, Open XML contains more identifiers, elements and explanatory examples and therefore its size is a lot bigger than ODF. Fast changes in IT field require this kind of standards to be published as fast as possible. For these reasons, like in the case of ODF, there is no problem with the fast standardization of Open XML.
- **This standard serves only to one company and not prepared with the cooperation of all related parties.** Many companies and organizations involved in the preparation process of this standard. Furthermore, this document group, with its target community, contains more than 90% of the documents in Turkey and all over the world. In the context of existing e-government and e-transformation Works, it is risky to transform these documents only by using ODF standard. The world doesn't comprise of only 'we and the other', there are many public and private organizations and people with different requirements not covered with this definition. Meanwhile, OpenXML is created, with the coordination of ECMA, with the representatives from Apple, Barclays Capital, BP, The British Library, Essilor, Intel, Microsoft, Next Page, Novell, Statoil, Toshiba and United States Library of Congress.
- **There are identifiers and settings with unknown content.** While there are IPR and patent related problems in both ODF and Open XML, it is not necessary for compliance to standard in a full implementation. Existence of these identifiers and settings for backward compatibility is a must and it's better to have them compared to not having them.



- **With backward compatibility claims, openness and interoperability compromised.** This is also related to backward compatibility with almost same rationale as in the previous clause. It is more meaningful to have these points in the standard and compromise the openness and interoperability.
- **Instead of using existing standards, Microsoft is forcing to use its own standards.** We observed the excessive use of ISO and other common standards in the Open XML specification. Additionally, as it is mentioned in two previous clauses, for backward compatibility reasons, some Microsoft and other source originated standards used in the specification. We don't think that they have to be used in a full implementation but they can be used to solve some problems.
- **It contradicts with other ISO standards and there are problems with its date format.** We don't think there are direct contradictions with ISO standards except the formats embedded to the specification to support the legacy systems.
- **Does Open XML independent from application, environment and operating system?** In this content, it is early to mention about a full independence for both ODF and Open XML. In order to specify this, IT sector must forget about the entire past document technologies. It has been observed that current works have targets about forward compatibility. Also, we see that it is theoretically impossible to combine these two standards into one standard. At the same time, we think that these two standards are two important milestones in achieving the real openness and interoperability that we hope will happen in the future.
- **Does Open XML really backward compatible?** While it is difficult to answer this question directly because there is no testing mechanism for this, when ODF and Open XML specifications are examined it is obvious that Open XML specification is stronger than ODF in backward compatibility.



4 Conclusion

After our research and work, the following results have been reached in addition to the above mentioned issues.

- Open XML addresses different requirements than ODF does.
- Since their purposes are the same but their targets are different, both ODF and Open may exist as two different standards. This is even a necessity.
- In the context of e-Government Works, the ODF by itself is not a solution for a large group of documents that are required to be converted. Although any loss of data is not expected in a direct way, the losses that may happen in formatting process may cause some data to be lost indirectly.
- Due to their design that makes interoperability possible in a flexible structure, Open XML and the other standards of similar kind, are thought to be necessary to provide full integrity of organisations in the e-Government Works of our country. For this reason, the direct contribution of our country to the development of Open XML and similar standards is considered to be important. Our participation to the voting process of ISO/IEC DIS 29500 is also evaluated to be a requirement.
- Due to Open XML related criticisms that have started to be discussed in a non-technical manner in time, as well as the discussions within the commercial and political framework; the support of our country to these standards is considerably important.

In addition to the technical issues above, considering the following non-technical issues of:

- the faster development rate in Computer Technology in comparison with the other areas, in spite of almost the same rate of consensus achieved; the clues which have been observed that the assessments of ODF and Open XML which have already turned to be a political way of discussions rather than technical ways and the possibility of continuation of these assessments in a deepening war";
- with the existence of ISO/IEC 26300, the expectations for ISO/IEC DIS 29500 to be a Standard, which make the parties closer than expected, prior to the long standing consensus expected to be realised in the years ahead;
- the conditions that the parties may move away from each other in case negative voting of OOXML for any reason

the ISO/IEC DIS 29500 is evaluated as **POSITIVE** to be accepted as a Standard and a positive country vote is recommended.