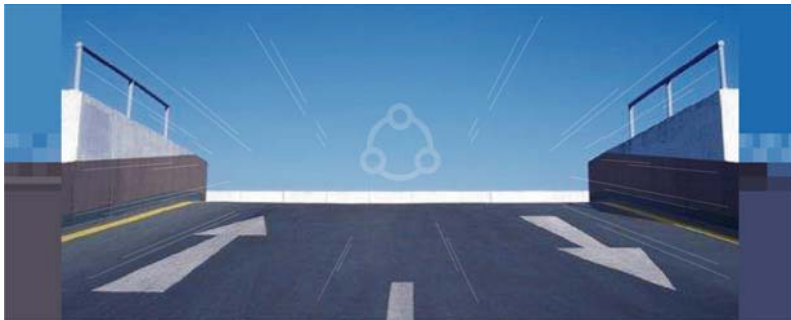


**Entrust**<sup>®</sup> Securing Digital Identities & Information



**Securing Your  
Digital Life**

***Forensic and Real-time Content Control***

Achieving Better Content Control in the Enterprise

February 2006

Entrust is a registered trademark of Entrust, Inc. in the United States and certain other countries. Entrust is a registered trademark of Entrust Limited in Canada. All other company and product names are trademarks or registered trademarks of their respective owners. The material provided in this document is for information purposes only. It is not intended to be advice. You should not act or abstain from acting based upon such information without first consulting a professional. ENTRUST DOES NOT WARRANT THE QUALITY, ACCURACY OR COMPLETENESS OF THE INFORMATION CONTAINED IN THIS ARTICLE. SUCH INFORMATION IS PROVIDED "AS IS" WITHOUT ANY REPRESENTATIONS AND/OR WARRANTIES OF ANY KIND, WHETHER EXPRESS, IMPLIED, STATUTORY, BY USAGE OF TRADE, OR OTHERWISE, AND ENTRUST SPECIFICALLY DISCLAIMS ANY AND ALL REPRESENTATIONS, AND/OR WARRANTIES OF MERCHANTABILITY, SATISFACTORY QUALITY, NON-INFRINGEMENT, OR FITNESS FOR A SPECIFIC PURPOSE.

© Copyright 2006 Entrust. All rights reserved.

## Table of Contents

1 Introduction .....	3
2 The Challenge .....	3
3 Underlying Content Analysis and Filtering Technology .....	5
4 Real-time Content Control .....	7
5 Forensic Content Control .....	9
6 The Entrust Solution .....	11
7 About Entrust.....	12

## 1 Introduction

Regulatory and corporate governance pressures are forcing organizations to scrutinize both real-time communications and those they store in their archives. Cases continue to be brought forth by the Securities Exchange Commission (SEC) that demonstrate the importance of monitoring, remediating and archiving key communications. Morgan Stanley<sup>1</sup> is being fined \$15 Million for not retaining key emails in relation to the Perleman case, which resulted in a \$1.54 Billion<sup>2</sup> ruling against the firm. Key to this ruling is that Morgan Stanley was not able to locate or find emails that it had archived for e-discovery.<sup>3</sup>

Information is being stored in the enterprise at astounding rates. IDC predicted<sup>4</sup> in Dec 2005 that the email archiving market will continue to grow by 34% annually. Terabytes of data are being stored by enterprises, creating a backlog of electronic information. Organizations need to sift through this information and keep what is required and delete what is not required. However, regulated organizations do not typically have a choice in terms of what they keep, and for this reason it is essential that they auto-classify incoming information to facilitate e-discovery.

The objective of this whitepaper is to describe solutions for both forensic and real-time content control. Effective content control is an end-to-end problem that requires solutions that can be integrated seamlessly in the enterprise—if content is sensitive and policy allows that it be sent, ensure that it is encrypted; if it is required for e-discovery, put it in the archive and make sure it is pre-tagged for easier classification. Look for vendors that offer as much of the solution as possible and who have partnerships with document and records management companies.

Entrust has a number of whitepapers on content control, including Content Analysis Approaches, Myths and Realities and Best Practices for Choosing a Content Control Solution, available for download at [www.entrust.com](http://www.entrust.com).

## 2 The Challenge

When discussing regulatory compliance and electronic communications, there are a multitude of legislative and regulatory issues to consider. For example, financial institutions face well over a dozen compliance regulations. An enterprise that is regulated may have obligations for examining any outbound content whether it is through email, instant messaging, file transfer or web postings, and this can vary by industry, creating a complex overall set of policy objectives. In response, many organizations have established compliance and risk management teams focused on electronic communications, with Compliance and Security Officers being asked to report to the CEO on any potential risks that may affect operations.

---

<sup>1</sup> <http://www.bloomberg.com/apps/news?pid=10000103&sid=aBoVvwOm0I6I&refer=us> Morgan Stanley to Pay \$15 Million for Failure to Save E-Mails, Bloomberg News, February 14/06.

<sup>2</sup> <http://www.allbusiness.com/periodicals/article/478745-1.html>

E-mail is key in judgment against Morgan Stanley. From Information Management Journal: July-August, 2005 issue.

<sup>3</sup> <http://www.washingtonpost.com/wp-dyn/content/article/2005/05/16/AR2005051601308.html> Perelman Wins Round in Suit Against Morgan Stanley, *Jill Barton*, Associated Press Tuesday, May 17, 2005; Page E03

<sup>4</sup> <http://www.idc.com/getdoc.jsp?containerId=prUS20047406> IDC Press Release, IDC Forecasts Continued Double-Digit Email Archiving Applications Market Growth through 2009. 17 January, 2006.

Depending on the industry sector, an organization may be dealing with a number of compliance issues, including:

- Public-company regulations, such as Sarbanes-Oxley, established in response to the Enron debacle;
- Regulations affecting financial services companies, such as banks and brokerages, who are required to adhere to Securities Exchange (SEC) rules, Graham-Leach Bliley (GLBA) and NASD or the National Association of Dealers;
- Regulations affecting healthcare privacy information, such as Health Insurance Portability and Accountability Act (HIPAA) which protects Personal Health Information (PHI);
- Intellectual property law, which is important for information asset protection in most organizations and particularly the Securities, Pharmaceutical, Technology and Manufacturing sectors;
- Regulations such as the Privacy Act and PIPEDA affecting the privacy of information, including personal identification information, such as PII information regularly collected from employees, customers and end users;
- Corporate Governance Policies, including disclosures to Boards of Directors and Auditors as well as Human Resources, Governance, Harassment and Code of Conduct and Ethics policies.

Often organizations break up the compliance requirements and solutions into two categories to address them: **Pre-review** and **Post-review**.

**Pre-review** implies that outgoing communications are monitored and action may be taken to block those that are non-compliant. Quarantining or forwarding of outgoing communication for review implies that one of the corporate policies or regulations has not been adhered to. Borderline policy non-compliance may result in a communication being audited or simply monitored and sent to the team within the organization that is responsible for compliance. Depending on the nature of the violation, this may be due to a regulatory issue requiring review by the Chief Counsel's office, an HR issue or some corporate governance issue. Relevant communications tools include real-time emails, Instant Messages, web postings or file transfers. The communication protocols for these are SMTP, IM, HTTP or FTP, respectively.

**Post-review** implies that the outgoing communication is archived or logged in a repository. In the case of email, this is often through an email archive. Email archive solutions have also expanded to include the archiving of Instant Messaging traffic and live news reports coming in over HTTP. Email archive solutions are converging with records and document management solutions. Records management solutions typically manage the lifecycle of a document from its point of creation through to its deletion, once the statute of limitation for that type of document expires. Document management solutions fulfill a similar function but also focus on the file plans of an organization and the classification of documents into the repository. A challenge for most email archive, records and document management solutions is the use of encryption. If an email or a document cannot be decrypted, that email or document cannot be viewed. Timely decryption capabilities are critical to complying with discovery and the audit demands of the Securities Exchange Commission.

### 3 Underlying Content Analysis and Filtering Technology

As a leader in security, Entrust acquired critical content analysis technology to help build a leading edge content control solution. The technology and intellectual property that Entrust acquired is a result of work conducted since the mid-1980s (in the area of Artificial Intelligence algorithms from government R&D labs) that was spun off into commercial products in 1998. As such, Entrust has experts in both security and content analysis and control, an important asset for the development of pre-review and post-review technology for compliance and e-discovery.

If one examines the underlying technology of typical compliance offerings as summarized in Table 1, some inadequate and outdated technologies will be identified. Many compliance and discovery tools used for search and tagging are still word-based. Word-based technologies were invented in the 1950s when computing was born. While these systems are being 're-discovered' in packet search approaches by packet-level compliance vendors who today, sell tools for monitoring Instant Messaging, Internet Mail and File Transfers and claim to do contextual analysis (where the context refers to the packet type), they produce the most false positives and are often inaccurate. However, they are most successful in exact matching if they can access database records. Should a credit card number not be contained within a database record, the exact matching capability of the packet system fails, and that credit card number will leave the enterprise without detection.

An innovation over word-based matching systems is rules-based technology which may be extended to use key phrases for matching. The rules-based technology forms the basis of most of the content filtering systems sold for email compliance solutions. Rules also offer advantages and disadvantages as shown in Table 1. Rules are very difficult to manage, especially as one considers that the English language has a subset of 20,000 words in daily use, out of a possible 1,000,000 in the dictionary. As such, for every word or set of words, a rule must be encoded with any exceptions to that rule. Rule-based systems became popular in the 1970s with the field of expert systems. Rule-based systems are difficult to administer, maintain and predict. Rule-based systems will alter their results unpredictably based on the order that rules are triggered as published by MIT researchers in the early 1980s. Furthermore, rule-based systems will also produce many false positives (although less than word-based systems) and present many false negative results that will burden system administrators, compliance officers and forensic auditors.

Many vendors have also used induction systems such as Bayesian or automatic learning algorithms to automatically derive rules through statistical analysis of texts and the assignment of probabilities to the presence or absence of words or key phrases and the building of word networks that link these words together. The problem with Bayesian approaches is that they are devoid of semantic or linguistic context. They are also typically fraught with false positives and introduce unnecessary complexity into compliance and forensic discovery solutions. These Bayesian approaches have to be re-trained on a regular basis adding to the administrative burden of the IT team.

Natural language processing (NLP) systems are commonly referred to as linguistic systems that typically break up a paragraph into a sentence and a sentence into its verb and noun phrase components, etc. NLP systems, while highly accurate, are very slow (as per Table 1) and are NP complete (meaning they may never return an answer within a finite time period unless that search is bound or stopped).

Method	Description of Approach	Advantages	Disadvantages
<b>Word Lists/Templates</b>	Compile list of <b>words</b> to detect in email/ document/ instant message/ web posting/etc.	Simple to understand	Time-consuming to create & maintain list. Quickly inaccurate.
<b>Rule Bases/Templates</b>	Compile <b>If-then</b> list of rules+ exceptions to detect in email / document/instant message/etc.	First set of 100 rules easy to compile.	As rule lists grow to tens of thousands and doubles with exceptions, lists become very hard to create/manage.
<b>Rule Bases thru Automatic Bayesian or Statistical Models</b>	Use <b>Machine Learning or ML</b> to generate probability patterns of words linked to words.	Automates the generation of Rule Sets.	Limited accuracy, high # false positives. Hard to tune complex model. Time consuming to update.
<b>Natural Language Processing (NLP)</b>	<b>NLP</b> breaks down a sentence into basic grammatical units of noun, verb, verb-phrase, etc.	Accuracy. Identifies the deep linguistic components.	Very slow, not suitable for thousands/millions of messages. Will not perform at wire speeds required for recognition of instant message traffic, etc.
<b>Concept Libraries/ Policy Modules: Pre-packaged &amp; Customizable</b>	Library of Concepts defined in a tree or hierarchical structure that more closely mimics how people relate information to other information. Can include full <b>ML</b> models of patterns representing specific content (e.g. intellectual property can be a concept leaf in the hierarchy of corporate policy). <b>Concepts follow object-oriented approach and can be re-used and inherit info.</b>	Consistent accuracy. Catches new patterns due to content analysis. Easily maintained due to object hierarchy. Supports remote updates. Uses <b>NLP &amp; statistical analysis</b> for best match.	Requires human expert knowledge for creation of concept library.  <b>NB: Entrust provides various concept library modules for Corporate Governance, Securities, SOX, Privacy, HIPAA, GLBA, etc. to greatly accelerate deployment.</b>

**Table 1: Content Analysis and Filtering Technologies**

The Entrust approach is based on a hybrid of automatic statistical and linguistic algorithm analysis of content as per Table 1. Statistical analysis allows the approach to be very fast (it has been proven through ISP and large customer deployments processing millions of messages in real-time) while bounded linguistic analysis allows it to be very accurate. Equally important is the fact that the patented concept library approach leverages the object-oriented paradigm. The libraries have inheritance and object encapsulation which tightly encodes patterns. The concept libraries can be used to link concepts together to build hierarchical relationships that represent workflow or business objects based on policies in an enterprise and the actions that are associated with the concepts being triggered. For example, a Corporate Compliance policy will have concepts that represent aggressive language hierarchically. These concepts may in turn inherit sub-concepts from offensive language policies. Furthermore, the aggressive language policy may inherit concepts and sub-concepts from coercive language and so on.

Identity as a high-level concept is another good example of the flexibility and power of the Entrust concept libraries. Identity is a concept with a number of sub-concepts relating to social security information, individual address information (home, email, phone numbers, etc.), education, racial

information, etc. The identity concept or object is present in the Privacy policy that examines documents for patterns that may be related to the privacy of employees, customers, and any persons associated with the enterprise. Identity is also a concept that is present in the securities policy as it relates to information about individuals and their dealings with an enterprise. Because of the object-oriented structures, the definition, administration and maintenance of the policy modules is far more manageable than pattern bases of Bayesian networks and rule-based systems.

Furthermore, the concept analysis capability also makes use of sophisticated matching techniques at the pattern level such as stemming (finding and using the root of words to disambiguate and speed up search), fuzzy matching, thresholds, soundex, misspellings, etc. The concepts can also leverage automatically derived models by learning a set of patterns from as little as 20 emails or documents. A good example of this is the policy for intellectual property. This policy has patterns that allow the automatic detection of contracts, source code, agreements, patents, etc. However, an organization may have patents associated with Biotechnology while another may have specific intellectual property associated with pharmaceutical drugs. A set of 20 emails or documents representing either type of IP can be used to learn a model that is then encapsulated into a concept for more customized matching. This allows the Entrust approach to be highly accurate and targeted. Entrust is unique in this approach and has over 12 recent content analysis patents granted and pending protecting this technology.

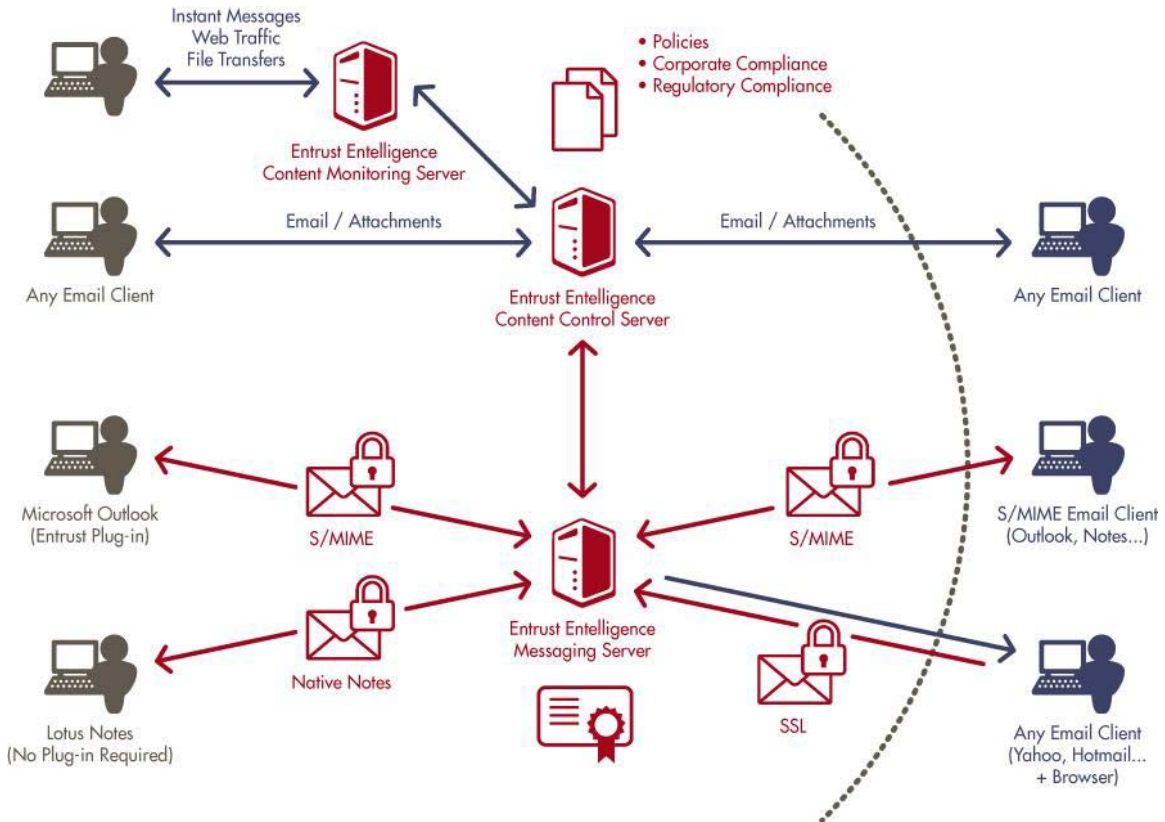
## 4 Real-time Content Control

### **Entrust Intelligence™ Content Control Server (ECCS), Entrust Intelligence™ Content Monitoring Server (ECMS), Entrust Intelligence™ Content Analysis Toolkit (ECAT) and Entrust Intelligence™ Messaging Server (EMS)**

The Entrust Intelligence™ Content Analysis toolkit bundles the unique content analysis technology of Entrust. This toolkit bundles the patented and highly sophisticated content analysis engine. The **Entrust Content Analysis Toolkit** has been in use by vendors and Entrust partners such as FileNet and Symantec in demanding applications such as Records and Document Management. End users use it to automatically generate a profile of a document and then have that document automatically and accurately categorized into a corporate taxonomy often referred to as an enterprise “file plan.” A major systems integrator recently integrated the toolkit with a typical platform for Records and Document Information Management. See [www.entrust.com](http://www.entrust.com) for more details on the toolkit.

**ECMS** monitors packet traffic for instant messages, file transfers, and web postings as shown in Figure 1. It routes content to ECCS for analysis. **ECCS** is a server-based product that bundles and leverages the content analysis toolkit to process millions of messages a day for Entrust customers. In the real-time mode, it acts as an SMTP gateway and automatically analyzes inbound and outbound email and attachments against a set of patterns based on the policies that the enterprise is running. As messages are analyzed or tagged, they can be blocked or quarantined, forwarded to a compliance officer, sent back to the user for reconsideration, etc. ECCS also analyzes any instant messages, file transfer or web traffic sent to it by ECMS. ECMS has an onboard recording and reporting database to which it writes the concepts or tags from the analyzed messages. This onboard database can be used to generate reports daily, weekly or on-demand. ECCS also has administrator and compliance officer consoles from which the reports can be examined. It is implemented using standard protocols and technologies and is fully scalable and architected for failsafe operations. It has been installed in environments processing millions of messages a day. ECMS and ECCS work together to remediate and take action on traffic types such as IM, FTP and HTTP (Internet Mail and Browsing) in real-time with all the rich functionality and actions that ECCS takes for email or SMTP traffic.

Email comes into the ECCS, is analyzed in real-time and is then tagged. If an email is outgoing and there is a policy that requires selective encryption, it is encrypted through EMS. Furthermore, an end user can be on an Outlook/Notes client or a Blackberry and sending a message with an attachment. That message is automatically scanned and tagged in real-time by ECCS. If it is found to be sensitive, it is encrypted through **EMS** before leaving the corporate network.



**Figure 1: Content Monitoring, Content Control and Selective Encryption**

## 5 Forensic Content Control

ECCS in a forensic mode can be used to automatically tag and analyze archive messages and documents as per Figure 2. This is accomplished through a connector to the email archive. As such, it can automatically analyze the emails and attachments and determine what concepts are reflected in them. Like the real-time ECCS, the forensic ECCS can analyze hundreds of attachment types in real-time. The concepts (or tags) are related to the encoded policies in the concept base. For example, the SEC module encapsulates patterns related to the securities rules, GLBA and NASD. Similarly, the SOX module encapsulates patterns related to the disclosures required in the Sarbanes-Oxley regulations on public company communications. These concepts can then be used as tags for the emails and attachments and used to better target search for e-discovery.

The Entrust Messaging Server (EMS) is a server that encrypts and decrypts email and attachments. It is used in conjunction with ECCS in real-time or in a forensic mode. In the real-time mode, ECCS is used to automatically scan content against corporate policy. If a message or document is found to be sensitive according to enterprise policy, it is sent by the ECCS to EMS to encrypt it in real-time. If a message or attachment is already encrypted, it is decrypted by EMS and then analyzed by ECCS. If it is found to contravene policy, it may be quarantined, sent to the compliance team for review or sent to the end user for reconsideration. Other actions for remediation include rejection (useful for offensive inbound messages), deletion, etc. These actions are attached to concepts that encapsulate patterns and trigger remediation actions according to enterprise policy, once they are detected in any inbound or outbound email or documents.

A typical forensic ECCS/EMS architecture is shown in Figure 2. Based on volumes, a typical application for a 50,000-user environment is estimated to require three ECCS servers to communicate with the email archive servers, as the speed of ECCS seems to be three times that of the typical archive. The fourth ECCS server is used to aggregate the tags into the database which also holds an index to the original email. Additional connectors to other litigation environments can also be developed.

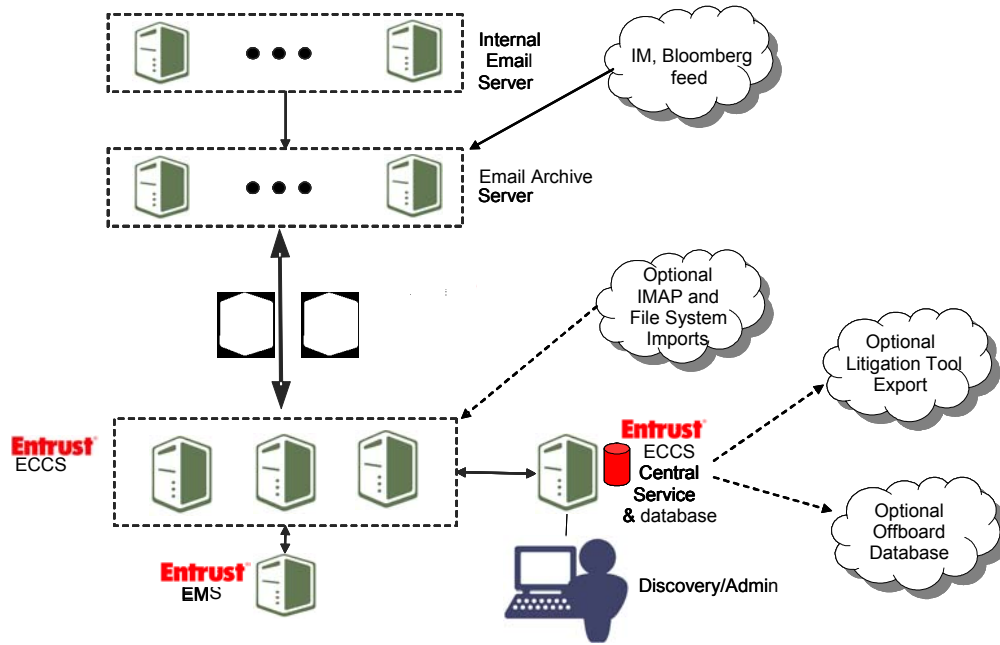


Figure 2: Forensic Tagging and Encryption/Decryption of Email Archives

## 6 The Entrust Solution

The Entrust Content Control and Secure Messaging Solution (as shown in Figure 1 above) offers a comprehensive solution with an integrated suite of components that can provide advanced content analysis of inbound and outbound messages, centralized policy enforcement, automatic and content-based email encryption, support for mobile devices and more. The solution can also be set up to monitor email, instant messaging, web traffic and file transfers in real-time. The capabilities have been designed for large enterprises and government organizations needing to enforce corporate or regulatory compliance and mitigate the risks of communicating sensitive information for thousands of users sending millions of messages each day.

As described, the Entrust Secure Messaging solution can also be used in forensic or real-time mode, assisting an organization in their e-discovery activities as well as offering a solution for immediate tagging of archives for discovery requirements and auditors.

Pre-defined or custom policies offer organizations the choice of subscribing to “plug-and-play” policy modules for: **Corporate Governance** (for protecting the privacy of customer and employee information, detecting harassment, offensive language, and protecting IP) and **Regulatory Compliance** (Sarbanes-Oxley, Securities Rules, NASD rules, Graham-Leach-Bliley – GLBA, Healthcare Portability and Accountability Act - HIPAA, etc.). Leveraging automatic enforcement of those policies—whether it be to block non-compliant communication, archive regulated information, bounce back emails with offensive language for reconsideration or automatically encrypt emails containing sensitive content or intellectual property—the solution does not rely on users to enforce policy and can provide a comprehensive set of capabilities that can be tailored for unique customer environments.

To learn more about the Entrust Solution for Content Control and Secure Messaging, please visit <http://www.entrust.com>.

## 7 About Entrust

Entrust, Inc. [NASDAQ: ENTU] is a world-leader in securing digital identities and information. Over 1500 enterprises and government agencies in more than 50 countries use Entrust solutions to help secure the digital lives of their citizens, customers, employees and partners. Our proven software and services help customers in achieving regulatory and corporate compliance, while helping to turn security challenges such as identity theft and e-mail security into business opportunities.